HIGH DIMENSIONAL VARIABLE SELECTION VIA PENALIZED LIKELIHOOD FOR GENERALIZED LINEAR MODELS

by

Wenjing Qi

B. S. in Statistics, Nankai University, China, 2008

Submitted to the Graduate Faculty of

the Kenneth P. Dietrich School of Arts and Sciences

in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in **Statistics**

University of Pittsburgh

2014

UNIVERSITY OF PITTSBURGH

KENNETH P. DIETRICH SCHOOL OF ARTS AND SCIENCES

This dissertation was presented

by

Wenjing Qi

It was defended on

August 25, 2014

and approved by

Satish Iyengar, PhD, Professor, Statistics

Leon J. Gleser, PhD, Professor, Statistics

Yu Cheng, PhD, Associate Professor, Statistics

Stewart J. Anderson, PhD, Professor, Biostatistics

Dissertation Director: Satish Iyengar, PhD, Professor, Statistics

Copyright © by Wenjing Qi
2014

HIGH DIMENSIONAL VARIABLE SELECTION VIA PENALIZED LIKELIHOOD FOR GENERALIZED LINEAR MODELS

Wenjing Qi, PhD

University of Pittsburgh, 2014

Variable selection is fundamental to high dimensional statistical modeling. In this study, penalized likelihood methods are examined to simultaneously estimate parameters and select variables for generalized linear models. We focus on the variable selection and parameter estimation properties rather than the prediction properties of the estimators and are more interested in situations where the number of parameters diverges with the sample size. We prove the parameter estimation consistency of several widely used penalized likelihood estimators for generalized linear models. We define the relaxed sense and prove that it loosens the regularity and sparsity conditions of the parameter estimation and variable selection consistency. We propose a bootstrap method that can greatly improve the variable selection performances and reduce false discovery rates. We conduct simulation studies to compare the variable selection and parameter estimation properties of these penalized likelihood estimators for logistic models. We then illustrate our methods on gene expression data.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1.0 INTRODUCTION

High dimensional data analysis has become increasingly common and important in diverse fields of sciences, engineering, and social sciences, with application to computational biology and health sciences to economics, financial engineering, risk management and machine learning. It characterizes many contemporary problems in statistics. For example, in disease classification using microarray data, tens of thousands of expressions of molecules are potential predictors while the number of tissue samples is usually less than 100. When interactions and higher order terms are considered, the dimensionality grows quickly. Other examples of high dimensional data include high-resolution images, high-frequency financial data, e-commerce data, warehouse data, functional data, longitudinal data, and text data, among others. To be more precise, high dimensionality here refers to the case where the dimensionality p is comparable to or larger than the sample size n.

High dimensional statistical problems suffer from statistical and computational challenges, called the "curse of dimensionality". One difficulty of high dimensional data analysis comes from the collinearity that often occurs among the predictors. That collinearity can easily be spurious in high dimensions (Fan and Lv, 2008), which can make us select a wrong model. Collinearity also gives rise to issues of overfitting, unstable parameter estimates and model misidentification. Overfitting is a major concern in high dimensional data analysis, and is usually blamed on high dimensionality. Avoiding overfitting is necessary to improve the generalization performance and the estimation accuracy. Another challenge of high dimensional data analysis is noise accumulation. Noise accumulation in high dimensional prediction has long been recognized in statistics and computer science. For example, classification using all variables can be as bad as a random guess due to noise accumulation in high dimensional space (Fan et al., 2008). They show that the difficulty of high dimensional classification is intrinsically caused by the existence of many noise variables that do not contribute to the reduction of classification error. In addition, the computational cost is very high for high dimensional statistical problems. Many traditional methods commonly used in

low dimensional data analysis are not even computationally feasible in high dimensional settings.

What makes high dimensional statistical inference possible is the assumption that the model lies in a low dimensional space. In such cases, the p-dimensional parameters are assumed to be sparse with many components being zero, where nonzero components indicate the important variables. Sparsity arises in many scientific problems. In genomic studies, it is generally believed that only a fraction of molecules are related to biological outcomes. For example, in disease classification, it is commonly believed that only tens of genes are responsible for a disease. Variable selection aims to identify all the important variables whose coefficients do not vanish and to provide effective estimates of those coefficients. With sparsity, variable selection can improve the generalization performance and the estimation accuracy, enhance the model interpretability and also help reduce the computational cost.

As pointed out in Fan and Li (2006), it is helpful to distinguish two types of statistical endeavors in high dimensional data analysis: prediction and parameter estimation. The former arises frequently in many statistical problems such as sales prediction and document classification. The latter appears naturally in many other contexts where we want to identify the significant variables and characterize the precise contribution of each to the response variable. Examples include health studies, where the relative importance of identified risk factors needs to be assessed for prognosis. Parsimonious models are desirable as they help to enhance the interpretation of the model, to gain insight into the relationship between predictors and response, and to provide a better understanding of the underlying process that generated the data.

Traditional variable selection procedures follow best subset selection and its stepwise variants. However, best subset selection is computationally too expensive for high dimensional problems. Furthermore, subset selection is unstable; thus, the resulting model has poor prediction accuracy and estimation accuracy. The classical AIC and BIC deal with the trade-off between the goodness of fit of the model and the complexity of the model. They can be regarded as L_0 penalized likelihood. The work of AIC and BIC suggests a unified approach to simultaneous parameter estimation and variable selection: penalized likelihood methods. Examples of widely used penalties are the L_1 penalty, the L_2 penalty, the $L_1 + L_2$ penalty, the L_q penalty ($0 < q \leq 2$), and the SCAD penalty, to mention only a few.

In this thesis we address the issues of variable selection and parameter estimation for high

dimensional statistical modeling in the unified framework of penalized likelihood methods for generalized linear models. The rest of the thesis is organized as follows. In Chapter 2, we introduce the penalized likelihood methods for generalized linear models. Chapter 3 reviews some recent advances in assessing the variable selection and parameter estimation properties of the penalized likelihood estimators. We prove parameter estimation consistency without sparsity assumptions of several popular penalized likelihood estimators for generalized linear models, define the relaxed sense, and present a simulation study of these penalized likelihood estimators for logistic models in Chapter 4. In Chapter 5, we discuss the consistency properties with sparsity assumptions of the penalized likelihood estimators for generalized linear models when p is much bigger than n and present several simulation studies. We propose a bootstrap method that can greatly improve the variable selection performances and reduce false discovery rates, and apply it to gene expression data in Chapter 6. In Chapter 7, we state future directions of study in this area.

2.0 PENALIZED LIKELIHOOD METHODS FOR GLMS

The data that we collect are usually of the form $(y_i, x_{i1}, \ldots, x_{ip})_{i=1}^n$, considered to be a random sample from some population. The random variable y is the response (or the dependent variable) and $\boldsymbol{x} = \{x_1, \ldots, x_p\}$ are the predictor variables (alternatively, the independent variables, the explanatory variables, or the covariates). These predictors may be the original measured variables and/or selected functions constructed from them. Generalized linear models (GLMs) provide a flexible parametric approach to estimating the covariate effects (MacCullagh and Nelder, 1989).

2.1 LIKELIHOOD OF GLMS

We denote by $\boldsymbol{X} = (\boldsymbol{x_1}, \ldots, \boldsymbol{x_p})$ the $n \times p$ design matrix with $\boldsymbol{x_j} = (x_{1j}, \ldots, x_{nj})^T, j = 1, \ldots, p$ and $\boldsymbol{y} = (y_1, \ldots, y_n)^T$ the n-dimensional response vector. With a canonical link, the conditional distribution of $\boldsymbol{y}$ given $\boldsymbol{X}$ belongs to the canonical exponential family with density

$$f_n(\boldsymbol{y}; \boldsymbol{X}, \boldsymbol{\beta}) \equiv \prod_{i=1}^{n} f_0(y_i; \theta_i) = \prod_{i=1}^{n} \{c(y_i) \exp[y_i\theta_i - b(\theta_i)]\} \tag{1}$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ is an unknown p-dimensional parameter vector, $\{f_0(y; \theta) : \theta \in \mathbb{R}\}$ is a regular exponential family, and $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)^T = \boldsymbol{X\beta}$.

In view of (1), the log-likelihood $\log f_n(\boldsymbol{y}; \boldsymbol{X}, \boldsymbol{\beta})$ of the sample is given, up to an affine transformation, by

$$l_n(\boldsymbol{\beta}) = \boldsymbol{y}^T\boldsymbol{X\beta} - \mathbf{1}^T\boldsymbol{b(X\beta)} = \sum_{i=1}^{n} \{y_i(\boldsymbol{x_i}^T\boldsymbol{\beta}) - b(\boldsymbol{x_i}^T\boldsymbol{\beta})\} \tag{2}$$

where $\boldsymbol{b}(\boldsymbol{\theta}) = (b(\theta_1), \ldots, b(\theta_n))^T$. For example, for logistic models, (2) becomes

$$l_n(\boldsymbol{\beta}) = \sum_{i=1}^{n} \{y_i(\boldsymbol{x_i}^T\boldsymbol{\beta}) - \log(1 + e^{\boldsymbol{x_i}^T\boldsymbol{\beta}})\};$$

in Poisson regression models, (2) can be written as

$$l_n(\boldsymbol{\beta}) = \sum_{i=1}^{n} \{y_i(\boldsymbol{x_i}^T\boldsymbol{\beta}) - \exp(\boldsymbol{x_i}^T\boldsymbol{\beta})\}.$$

2.2 PENALIZED LIKELIHOOD METHODS AND VARIABLE SELECTION

It is well known that the $\hat{\boldsymbol{\beta}}$ that maximizes the log-likelihood $l_n(\boldsymbol{\beta})$ of the sample (the MLE) often provides a poor estimate of the true parameter vector $\boldsymbol{\beta}^*$, because the mean squared error, MSE($\hat{\boldsymbol{\beta}}^{\boldsymbol{MLE}}$), is large. This is especially the case in high dimensional settings. It is caused by the high variability of the estimates $\hat{\boldsymbol{\beta}}$ when the log-likelihood is evaluated on different random samples drawn from the population distribution. Subset selection, the standard technique for improving the MLE, provides interpretable models but can be extremely unstable because it is a discrete process — predictors are either retained or dropped from the model. Small changes in the data can result in very different models being selected. Also it is computationally too costly for high dimensional problems. Moreover, it is well known that the MLE gives nonzero estimates to all coefficients, thus does not do variable selection and, in fact, is not unique when $p > n$.

A common remedy is to modify the log-likelihood in order to stabilize the estimates by adding a penalty on the parameter values. By doing so, we sacrifice by adding a little bias to reduce the variance of the estimates. Penalized likelihood methods usually involve automatic variable selection and are computationally feasible.

We consider maximizing the following penalized log-likelihood,

$$Q_n(\boldsymbol{\beta}) = l_n(\boldsymbol{\beta}) - \lambda P(\boldsymbol{\beta}), \tag{3}$$

where $P(\cdot)$ is a nonnegative penalty function of the parameters and $\lambda \geq 0$ is the parameter that regulates the strength of the penalty. Setting $\lambda = \infty$ produces the totally constrained solution whereas $\lambda = 0$ yields the unrestricted solution. Intermediate values $0 < \lambda < \infty$ provide degrees of

restriction between these two extremes, thereby regulating the stability (variance) of the estimates with respect to different training samples drawn from the population distribution. Thus maximizing (3) produces a family of estimates in which each member of the family is indexed by a particular value of λ:

$$\hat{\boldsymbol{\beta}}(\lambda) = \arg\max_{\boldsymbol{\beta}} \ Q_n(\boldsymbol{\beta}) = \arg\max_{\boldsymbol{\beta}} \ l_n(\boldsymbol{\beta}) - \lambda P(\boldsymbol{\beta}) = \arg\min_{\boldsymbol{\beta}} \ -l_n(\boldsymbol{\beta}) + \lambda P(\boldsymbol{\beta}). \tag{4}$$

This family lies on a one-dimensional path in the p-dimensional space of all joint parameter values. The same family of solutions can be obtained through the equivalent restriction form

$$\hat{\boldsymbol{\beta}}(t) = \arg\max_{\boldsymbol{\beta}} \ l_n(\boldsymbol{\beta}) \qquad \text{such that } P(\boldsymbol{\beta}) \leq t. \tag{5}$$

Penalized likelihood methods are continuous processes that shrink coefficients and hence are more stable. The shrinkage and the complexity of the model is determined by the penalty parameter λ. A large value of λ tends to choose a simple model, whereas a small value of λ inclines to a complex model. The estimation using a larger value of λ tends to have smaller variance, whereas the estimation using a smaller value of λ inclines to smaller modeling biases. The trade-off between the biases and variances yields an optimal choice of penalty parameter. Choosing the penalty parameter is an important part in penalized likelihood methods. It is often done by cross-validation.

The penalized log-likelihood can also be interpreted as the posterior log density with the prior density of the parameters proportional to $\exp(-\lambda P(\boldsymbol{\beta}))$. For example, the L_2 penalized likelihood (the ridge) can be regarded as having a Gaussian prior, and the L_1 penalized likelihood (the Lasso) can be regarded as having a Laplace prior.

Penalized regression (or penalized least squares) is a special case of penalized likelihood methods. When Gaussian distribution is the underlying distribution, the negative log-likelihood $-l_n(\boldsymbol{\beta})$ is proportional to the least squares $\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2$, and the penalized likelihood estimator is reduced to the penalized least squares estimator,

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \ \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 + \lambda P(\boldsymbol{\beta}) \tag{6}$$

Actually, the penalized log-likelihood itself is a special case of the penalized loss, as the negative log-likelihood function can be viewed as a loss function. Even more generally, the likelihood does not have to be the true likelihood. It can be a quasi-likelihood or a partial likelihood so that it

can be applied to models like the Cox's proportional hazards model.

2.3 SPARSITY

One property of the true parameter vector $\boldsymbol{\beta}^*$ that is often suspected is sparsity. That is, only a small fraction of the predictor variables $\{x_j\}_1^n$ are influencial predictors, and the majority of $\boldsymbol{\beta}^* = (\beta_0^*, \ldots, \beta_p^*)^T$ are exactly zero. Sparsity is an important feature of high dimensional models; statistical inference would not be possible without the assumption of sparsity because of statistical and computational difficulties. It simultaneously allows us to efficiently learn a model and to efficiently predict responses of new samples.

Without loss of generality, assume that $\boldsymbol{\beta}^* = (\boldsymbol{\beta}_\mathbf{1}^{*T}, \boldsymbol{\beta}_\mathbf{0}^{*T})^T$ with each component of $\boldsymbol{\beta}_\mathbf{1}^*$ nonzero and $\boldsymbol{\beta}_\mathbf{0}^* = \mathbf{0}$. We refer to the support (sometimes called the active set) $\mathrm{supp}(\boldsymbol{\beta}^*) = \{1, \ldots, s\}$ as the true underlying sparse model of the indices. Next, write $\boldsymbol{X} = (\boldsymbol{X}_\mathbf{1}, \boldsymbol{X}_\mathbf{0})$, where $\boldsymbol{X}_\mathbf{1}$ and $\boldsymbol{X}_\mathbf{0}$ are the first s and the last $(p-s)$ columns of $\boldsymbol{X}$, corresponding to $\boldsymbol{\beta}_\mathbf{1}^*$ and $\boldsymbol{\beta}_\mathbf{0}^*$ respectively. Variable selection aims at locating those predictors x_j with nonzero β_j^* and giving an efficient estimate of $\boldsymbol{\beta}_\mathbf{1}^*$. The degree of sparsity of the true parameter vector $\boldsymbol{\beta}^*$ can be defined as $1 - s/p$.

3.0 PROPERTIES OF PENALIZED LIKELIHOOD ESTIMATORS

3.1 PENALIZED LEAST SQUARES ESTIMATORS

A large number of studies has been done to introduce penalties and to assess the properties of penalized likelihood estimators for least squares problems.

3.1.1 The introduction of L_q penalties

A natural generalization of the L_0 penalized likelihood is the L_q penalized likelihood. The L_q penalized likelihood method is called the bridge regression in Frank and Friedman (1993), in which $P(\boldsymbol{\beta}) = \sum_{j=1}^{p} |\beta_j|^q$ for $0 < q \leq 2$. It includes the ridge ($q = 2$) and the Lasso ($q = 1$) as special cases. It is known that if $0 < q \leq 1$, bridge estimators produce sparse models thus does variable selection. If $1 \leq q \leq 2$, the L_q penalty is convex thus the optimization of the L_q penalized likelihood is a convex problem which is very attractive for computational purposes.

The L_2 penalty $P(\boldsymbol{\beta}) = \|\boldsymbol{\beta}\|_2^2 = \sum_{j=1}^{p} {\beta_j}^2$ has nice algorithmic and numerical features. Hoerl and Kennard (1970) propose the ridge estimator, which is the L_2 penalized likelihood estimator. The ridge estimator has frequently better performance than the MLE. However, it does not select variables. It does not force any parameter to be zero; hence it is unable to produce a parsimonious model for high dimensional problems.

The Lasso (Tibshirani, 1996) is a popular method that uses the L_1 penalty $P(\boldsymbol{\beta}) = |\boldsymbol{\beta}| = \sum_{j=1}^{p} |\beta_j|$ to achieve a sparse solution, thus it does automatic variable selection. It is widely used because the L_1 penalty is the only convex penalty that does automatic variable selection and also

because there is an efficient and statistically motivated algorithm called LARS (Efron et al., 2004) to implement it.

Although the Lasso is useful in many situations, it has some limitations. First, in the $p > n$ case, the Lasso selects at most n variables before it saturates, because of the nature of the convex optimization problem. This seems to be a limiting feature for a variable selection method. Second, collinearity can severely degrade the performance of the Lasso. The collinearity problem is often encountered in high-dimensional data analysis. Even when the predictors are independent, as long as the dimension is high, the maximum sample correlation can be large, as shown in Fan and Lv (2008). For the usual $n > p$ situations, if there are high correlations between predictors, it has been empirically observed that the performance of the Lasso is dominated by the ridge.

3.1.2 The elastic net and the grouping effect

Zou and Hastie (2005) propose the elastic net penalty, which is a compromise between the L_2 penalty ($\alpha = 1$) and the L_1 penalty ($\alpha = 0$),

$$P_\alpha(\boldsymbol{\beta}) = \alpha \|\boldsymbol{\beta}\|_2^2 + (1-\alpha)|\boldsymbol{\beta}| \tag{7}$$

where $0 \leq \alpha \leq 1$. This penalty is particularly useful in the $p > n$ situations, or any situation where there are many correlated predictor variables.

The ridge estimator is known to shrink the coefficients of correlated predictors towards zero and each other, allowing them to borrow strength from each other, and preventing the model from being poorly determined and exhibiting high variance. Usually it achieves a stable fit even in the presence of highly correlated predictors. In the extreme case of k identical predictors, they each get identical coefficients with $\frac{1}{k}th$ the size that any single one would get if fit alone. From a Bayesian point of view, the ridge penalty is ideal if there are many predictors, and all have non-zero coefficients (drawn from a Gaussian distribution).

The Lasso, on the other hand, is somewhat indifferent to very correlated predictors, and will tend to pick one and ignore the rest. As shown in Zou and Hastie (2005), the Lasso solution paths are unstable when predictors are highly correlated. In the extreme case above, the Lasso problem breaks down. The Lasso penalty corresponds to a Laplace prior, which expects many coefficients

to be close to zero, and a small subset to be larger and nonzero.

The elastic net with $\alpha = \epsilon$ for some small $\epsilon > 0$ performs much like the Lasso, but removes any degeneracies and wild behavior caused by extreme correlations. More generally, the entire family P_α creates a useful compromise between the ridge and the Lasso. As α decreases from 1 to 0, for a given λ the sparsity increases monotonically from 0 to the sparsity of the Lasso solution. Zou and Hastie (2005) give a Bayesian interpretation of the elastic net penalty as a mixture of Gaussian and Laplace prior. The parameter α determines the mix of the penalties. The elastic net can select more than n variables when $p > n$, another potential advantage over the Lasso.

The L_1 part of the elastic net performs automatic variable selection, while the L_2 part stabilizes the solution paths and, hence, improves the performance. In an orthogonal design where the Lasso is shown to be optimal, the elastic net automatically reduces to the Lasso. However, when the correlations among the predictors become high, the elastic net can significantly improve the estimation accuracy of the Lasso.

The grouping effect is a desirable property of a variable selection method in high dimensional problems. Qualitatively speaking, a method exhibits the grouping effect if the estimated parameters of highly correlated variables tend to be equal (up to a change of sign if negatively correlated). In particular, in the extreme situation where some variables are exactly identical, the method should assign identical coefficients to the identical variables. Zou and Hastie (2005) give a lemma that shows a clear distinction between strictly convex penalty functions and the Lasso penalty in the least squares problems.

Zou and Hastie's Lemma. *Consider penalized least squares problem* (6)*. Assume that* $\boldsymbol{x_i} = \boldsymbol{x_j}, i, j \in \{1, \ldots, p\}$.
(a) If $P(\cdot)$ *is strictly convex, then* $\hat{\beta}_i = \hat{\beta}_j, \forall \lambda > 0$.
(b) If $P(\boldsymbol{\beta}) = |\boldsymbol{\beta}|$*, then* $\hat{\beta}_i \hat{\beta}_j \geq 0$ *and* $\hat{\boldsymbol{\beta}}'$ *is another minimizer of* (6)*, where*

$$\hat{\beta}'_k = \begin{cases} \hat{\beta}_k & \text{if } k \neq i \text{ and } k \neq j, \\ (\hat{\beta}_i + \hat{\beta}_j) \cdot (s) & \text{if } k = i, \\ (\hat{\beta}_i + \hat{\beta}_j) \cdot (1 - s) & \text{if } k = j. \end{cases}$$

for any $s \in [0, 1]$.

Convexity guarantees the grouping effect. In contrast, the Lasso does not even have a unique solution. The elastic net penalty with $\alpha > 0$ is strictly convex, thus enjoying the property that guarantees the grouping effect in the extreme situation with identical predictors.

3.1.3 Lasso extensions

In recent years, there has been a huge amount of research activity devoted to Lasso extensions such as the fused Lasso (Tibshirani et al., 2005) and the group Lasso (Yuan and Lin, 2006). The fused Lasso is a generalization that is designed for problems with predictors that can be ordered in some meaningful way. The fused Lasso penalizes the L_1-norm of both the coefficients and their successive differences. Thus it encourages sparsity of the coefficients and also sparsity of their differences — i.e. local constancy of the coefficient profile. The group Lasso penalty is intermediate between the L_1 penalty and the L_2 penalty. It is L_2 among variables in the same groups and L_1 across different groups. The group Lasso includes or excludes variables in groups. It is very useful when there are predefined clusters of predictors like dummy coding categorical variables.

3.1.4 The sampling properties of interest

The sampling properties of the penalized likelihood estimators have been extensively studied for least squares problems. For the purpose of variable selection, we are concerned with the sparsity of the estimator, particularly its variable selection consistency meaning that the estimator $\hat{\boldsymbol{\beta}}$ has the same support as the true parameter vector $\boldsymbol{\beta}^*$ with asymptotic probability one, $P(\mathrm{supp}(\hat{\boldsymbol{\beta}}) = \mathrm{supp}(\boldsymbol{\beta}^*)) \to 1$ as $n \to \infty$. Zhao and Yu (2006) characterize the variable selection consistency by studying a scientifically meaningful and technically convenient property of sign consistency: $P(\mathrm{sgn}(\hat{\boldsymbol{\beta}}) = \mathrm{sgn}(\boldsymbol{\beta}^*)) \to 1$ as $n \to \infty$. The parameter estimation properties are also of great interest. The parameter estimation consistency and the asymptotic normality of the parameter estimates are two properties that we focus on. Fan and Li (2001) propose the oracle property, meaning that a method is variable selection consistent and that the estimators of the nonzero coefficients are asymptotically normal with the same means and covariance that they would have if the nonzero coefficients were known in advance.

3.1.5 The sampling properties in fixed p case

There is a large literature devoted to studying the variable selection and parameter estimation properties of the penalized least squares assuming the dimensionality p is fixed and the sample size n goes to infinity. The true parameter vector $\boldsymbol{\beta}^*$ is assumed to be fixed in this case.

Several authors have studied the properties of the Lasso. Knight and Fu (2000) show that, under appropriate conditions, the Lasso is consistent for estimating the regression coefficients and its limiting distribution can have positive probability mass at zero when the true value of the parameter is zero. However, careful inspection of their results indicates that the positive probability mass at zero is less than one in the limit for certain configurations of the parameters, which suggests that the Lasso is not variable selection consistent without additional assumptions. Leng et al. (2006) showed that the Lasso is in general not path consistent in the sense that (1) with probability greater than zero, the whole Lasso path may not contain the true parameter value; and (2) even if the true parameter value is contained in the Lasso path, it cannot be achieved by using prediction accuracy as the selection criterion. Zou (2006) further studied the variable selection and parameter estimation properties of the Lasso. He shows that the positive probability mass at zero of the Lasso, when the true value of the parameter is zero, is in general less than one, which implies that the Lasso is in general not variable selection consistent. He also provides a condition on the design matrix for the Lasso to be variable selection consistent. This condition was also discovered by Meinshausen and Bühlmann (2006) and Zhao and Yu (2006). In particular, Zhao and Yu (2006) call this condition the irrepresentable condition on the design matrix: there exists a positive constant vector $\boldsymbol{\eta}$

$$\|(\boldsymbol{X_1^T X_1})^{-1}\boldsymbol{X_1^T X_0}\| \leq \mathbf{1} - \boldsymbol{\eta},$$

where $\mathbf{1}$ is a $p - s$ by 1 vector of 1's and the inequality holds element-wise. However, the irrepresentable condition can easily be violated and become restrictive in high dimensions. This demonstrates that in high dimension problems, the Lasso estimator can easily select an inconsistent model. Therefore, the Lasso is variable selection consistent under certain conditions, but not in general. Further, the value of the penalty parameter λ required for variable selection consistency overshrinks the nonzero coefficients, which leads to asymptotically biased estimates of the nonzero coefficients. Hence, if the Lasso is variable selection consistent, then it is not consistent for estimating the nonzero parameters. Therefore, these studies confirm the conjecture that Lasso does not possess the oracle property (Fan and Li, 2001).

The bias of the Lasso estimator makes the Lasso prefer a smaller λ. As a result in model selection, when λ is automatically selected by a data-driven rule to compensate the bias problem, the Lasso estimator has to choose a smaller λ in order to have a desired mean squared error. Yet, a smaller value of λ results in a complex model. This explains why the Lasso estimator tends to have many false positive variables in the selected model. Therefore, the Lasso is good at finding (asymptotically) a superset of the correct predictors, and that methods that produce even sparser models can be useful.

Zou (2006) proposes the adaptive Lasso, where adaptive weights are used for penalizing different coefficients in the L_1 penalty,

$$P(\boldsymbol{\beta}) = \sum_{j=1}^{p} \hat{w}_j |\beta_j| \tag{8}$$

The adaptive weights are determined by an initial estimator, that is, $\hat{w}_j = |\hat{\beta}_j^{initial}|^{-\gamma}$, where $\gamma > 0$. The adaptive weights will reduce the penalty for large parameters and thus reduce the bias for large parameter estimates. Therefore, the adaptive Lasso is a promising approach for producing even sparser models than the Lasso. He also shows that, if a reasonable initial estimator is available, then under appropriate conditions, the adaptive Lasso enjoys the oracle property.

The adaptive Lasso penalty is not strictly convex hence like the Lasso it does not even have a unique solution in the extreme situation with identical predictors from Zou and Hastie's Lemma. The adaptive Lasso does not have the grouping effect. It inherits the instability of the Lasso for high dimensional data.

On the other hand, the elastic net is not an oracle procedure even for the usual $p < n$ case. Notably to solve the original elastic net problem, Zou and Hastie (2005) transform the elastic net into ordinary Lasso type problem in some augmented space by some algebraic manipulation. Since this is a one-to-one mapping, whenever the Lasso is inconsistent in the augmented space, so is the underlying elastic net. Yuan and Lin (2007) state an explicit condition for the inconsistency of the elastic net similar to the irrepresentable condition for the Lasso.

The adaptively weighted L_1 penalty and the elastic net penalty improve the Lasso in two different directions. The adaptive Lasso achieves the oracle property and the elastic net handles the collinearity. Zou and Zhang (2009) propose the adaptive elastic net that uses a combination of the

adaptive L_1 penalty and the L_2 penalty, thus achieve the two improvements simultaneously. The adaptive elastic net penalty is given by

$$\lambda P(\boldsymbol{\beta}) = \lambda_1^* \sum_{j=1}^{p} \hat{w}_j |\beta_j| + \lambda_2^{enet} \|\boldsymbol{\beta}\|_2^2 \tag{9}$$

where $\lambda_2^{enet} = \alpha^{enet}\lambda^{enet}$ is the L_2 penalty parameter in the elastic net estimator and the adaptive weights are constructed by $\hat{w}_j = |\hat{\beta}_j^{enet}|^{-\gamma}$, $\gamma > 0$. The adaptive elastic net is shown to have the oracle property under weak regularity conditions. And the adaptive elastic net penalty is strictly convex (for $\lambda^{enet} > 0$) hence it enjoys the unique solution in the extreme situation with identical predictors from Zou and Hastie's Lemma. Thus, the adaptive elastic net has the grouping effect.

Fan and Li (2001) advocate penalty functions that result in an estimator with three properties:

1. Unbiasedness: The resulting estimator is nearly unbiased when the true unknown parameter is large to avoid unnecessary modeling bias.

2. Sparsity: The resulting estimator is a thresholding rule, which automatically sets small estimated coefficients to zero to reduce model complexity.

3. Continuity: The resulting estimator is continuous in data to avoid instability in model estimation.

In general for the penalty function, the singularity at the origin is needed for generating sparsity in variable selection and the concavity is needed to reduce the estimation bias. In fact, the L_q penalty with $0 < q < 1$ does not satisfy the continuity condition, the L_1 penalty does not satisfy the unbiasedness condition, and the L_q penalty with $q > 1$ does not satisfy the sparsity condition and the unbiasedness condition. Therefore, none of the L_q penalties satisfies the above three conditions simultaneously, and the L_1 penalty is the only penalty that is both convex and produces sparse solutions but it does not have the oracle property. Huang et al. (2008a) show that the bridge estimator with $0 < q < 1$ has the oracle property under appropriate conditions. However, it does not have the continuity property.

The penalty functions satisfying the aforementioned three conditions are infinitely many. Fan and Li (2001) propose a non-concave penalty function referred as the smoothly clipped absolute deviation (SCAD). It corresponds to a quadratic spline function with knots at λ and $a\lambda$. This penalty leaves large values of parameters not excessively penalized and makes the solution continuous. It is given by $P(\boldsymbol{\beta}) = \sum_{j=1}^{p} p(|\beta_j|)$ with

$$p'(|\beta_j|) = n\{I(|\beta_j| \leq \lambda) + \frac{(a\lambda - |\beta_j|)_+}{(a-1)\lambda} I(|\beta_j| > \lambda)\} \tag{10}$$

for some $a > 2$.

The SCAD enjoys the oracle property, however, as a method using a nonconcave penalty, the objective function with the SCAD is not convex, so it is more difficult to compute. There have been efforts to devise efficient algorithms for non-convex penalized problems. For example, Fan and Li (2001) propose the local quadratic approximation (LQA) algorithm.

3.1.6 The sampling properties in diverging p_n case

The asymptotic theory with $p = p_n \to \infty$ seems to be more applicable to problems involving a huge number of predictors, such as microarray analysis and document/image classification. However, there are relatively few studies on the variable selection and parameter estimation properties of the penalized likelihood estimators with a diverging number of parameters. The true parameter vector $\boldsymbol{\beta}^*$ is assumed to change with the sample size n in this case.

Fan and Peng (2004) prove the $\sqrt{n/p_n}$ consistency and the oracle properties of the SCAD estimator in the moderate dimensional setting with $p_n = o(n^{1/5})$ under some regularity conditions. Zou and Zhang (2009) show that the adaptive elastic net estimator is $\sqrt{n/p_n}$ consistent and has the oracle property with $p_n = O(n^\nu)$ for some $0 \leq \nu < 1$ under some regularity conditions. Huang et al. (2008b) show that the adaptive Lasso has the oracle property even when $p_n > n$ with an appropriate initial estimator under some regularity conditions. It largely remains to show the consistency properties of the penalized likelihood estimators especially in the setup that p_n grows faster than n.

In studies on variable selection and parameter estimation consistency, the nonsparsity size s is rarely mentioned. The nonsparsity size s should be allowed to grow with the sample size n in order to be applicable in high dimensional settings and should be denoted s_n. In fact, the nonzero part of the true parameter vector $\boldsymbol{\beta}^*$, $\boldsymbol{\beta}_1^*$, should be allowed to change with the sample size n. Zhao and Yu (2006) show that $s_n = O(n^c)$ for some $0 \leq c < 1$ is needed for the Lasso to be consistent for variable selection. Fan and Lv (2011) prove that for nonconcave penalties, like the Lasso and the SCAD penalties, the penalized likelihood estimators are $\sqrt{n/s_n}$ consistent and have the oracle

property with $\log p_n = O(n^a)$ for some $a \in (0, 1)$ and $s_n = o(n^{1/3})$ under some regularity conditions.

3.2 PENALIZED LIKELIHOOD ESTIMATORS FOR GLMS

There is not much literature that studies the variable selection and parameter estimation consistency of the penalized likelihood estimators in the context of the GLMs. Fan and Li (2001) briefly discuss the regularity conditions for the SCAD estimator to have parameter estimation consistency and the oracle property for the GLMs when p is fixed. Zou (2006) show that the oracle property still holds for the adaptive Lasso estimator under mild regularity conditions for the GLMs when p is fixed. Fan and Peng (2004) give regularity conditions for the SCAD estimator to have parameter estimation consistency and the oracle property for the GLMs when p_n diverges at a much slower rate than n. Fan and Lv (2011) prove that under some regularity conditions for Lasso and SCAD penalties, the penalized likelihood estimators are parameter estimation consistent and have the oracle property even when p_n grows exponentially fast comparing to n, given that the true parameter vector is very sparse. It is still largely an open problem to show the variable selection and parameter estimation properties of the other penalized likelihood estimators for fixed p and diverging p_n cases in the context of the GLMs.

3.3 IMPLEMENTATION OF PENALIZED LIKELIHOOD ESTIMATORS

Efron et al. (2004) develop an efficient algorithm called the least angle regression (LARS) for computing the entire regularization path for the Lasso for linear regression models. Their algorithm exploits the fact that the coefficient profiles are piecewise linear, which leads to an algorithm with the same computational cost as the full least squares fit on the data. Many Lasso related procedures such as the adaptive Lasso and the elastic net can be transformed to the Lasso, thus can be implemented by the LARS algorithm. Rosset and Zhu (2007) characterize the class of penalized loss problems where piecewise-linearity exists — both the loss function and the penalty have to be quadratic or piecewise linear.

The solution path of the penalized likelihood of GLMs is piecewise smooth rather than piece-

wise linear, because the negative log-likelihood is not piecewise quadratic. Exact methods are slower than the LARS algorithm, and are less feasible when p is large. Many algorithms for GLMs are based on local quadratic approximations to the log-likelihood in the neighborhood of the current parameter estimates. These quadratic approximations generate iteratively reweighted least squares (IRLS) sub-problems that can be solved using simpler methods. Friedman et al. (2010) use cyclical coordinate descent algorithms, computed along a regularization path, for the elastic net and related convex penalties. Coordinate descent algorithms are extremely simple and fast, with an explicit formula for each coordinate-wise optimization, and exploit the assumed sparsity of the model to great advantage.

The L_q $(0 < q < 1)$, and the SCAD penalty functions are singular at the origin, and they do not have continuous second order derivatives. Therefore, computing their corresponding estimators is challenging. Fan and Li (2001) propose a generic and effective local quadratic approximation (LQA) algorithm for optimizing nonconcave penalized likelihood. Their idea is to locally approximate the objective function by a quadratic function. The penalty function can be locally approximated by a quadratic function and the log-likelihood can be locally approximated by a quadratic function as well for GLMs. Hence, maximizing the penalized likelihood becomes a least squares problem that admits a closed-form solution. They also show that LQA can provide a sandwich formula for computing the covariance of the estimates $\hat{\boldsymbol{\beta}}_{\mathbf{1}}$, the nonzero component of $\hat{\boldsymbol{\beta}}$.

When the dimensionality p_n is beyond the scale that a method can handle, a natural idea is to reduce p_n from a huge scale (say, $\log p_n = O(n^a)$ for some $a > 0$) to a relatively large scale (e.g., $p_n = O(n^b)$ for some $b > 0$). This provides a reduction in the number of variables that need to be entered into the optimization. Sure independence screening (SIS), introduced by Fan and Lv (2008), is capable of retaining all the important variables with asymptotic probability one. Tibshirani et al. (2012) also propose sequential strong rules to discard predictors.

3.4 PRACTICAL ISSUES

It is important to choose the penalty parameter λ that balances variance and bias. This is usually done by cross-validation in studies on penalized likelihood. However, there is no consensus on how to choose the fraction of samples reserved for training and for validation and the estimator may be prone to overfitting the cross-validation error (Cawley and Talbot, 2010). Generalized

cross-validation and AIC/BIC are sometimes used to determine the penalty parameter. But they are empirically observed to include unimportant variables in the selected model. Moreover, Leng et al. (2006) show in general that prediction-based methods for tuning are not sufficient for the Lasso and related methods in problems where the primary goal is selecting the set of true variables. Tuning for prediction rather than variable selection tends to create nonzero coefficients when the true coefficient is zero, but not the other way around. Methods of choosing the penalty parameter need to be explored for variable selection purpose.

Many regularity conditions for variable selection and parameter estimation consistency involve the convergence rate of the penalty parameter λ (or more precisely λ_n) as n goes to infinity. However, in practice, the penalty parameter is determined by data-driven methods. The variable selection and parameter estimation consistency properties need to be established for situations when the penalty parameter is chosen by data-driven methods.

Leeb and Pötscher (2005, 2008) criticize the oracle property in the context of penalized regression for holding only pointwise in the parameter space and giving a misleading picture of the actual finite sample performance of the estimator. They argue that the finite sample properties of an estimator enjoying the oracle property are often markedly different from what the pointwise asymptotic theory predicts. The finite sample distribution can be bimodal regardless of sample size, although the pointwise asymptotic distribution is normal. The finite sample distribution can escape to infinity along appropriate local alternatives although the pointwise asymptotic distribution is perfectly normal. More studies are needed on the finite sample behavior of the penalized likelihood estimators that have nice pointwise asymptotic properties.

4.0 CONSISTENCY WITHOUT SPARSITY ASSUMPTIONS

4.1 THE GROUPING EFFECT

We prove a lemma on the grouping effect for GLMs as a generalization of Zou and Hastie's Lemma. It shows the same clear distinction between strictly convex penalty functions and the Lasso penalty for GLMs. Strictly convex penalties like the L_2 penalty will be needed to guarantee the grouping effect in high dimensional problems.

Lemma 1. *Consider penalized likelihood estimator* (4)*. Assume that* $\boldsymbol{x_i} = \boldsymbol{x_j}$*, for some* $i, j \in \{1, \ldots, p\}$.

(a) If $P(\cdot)$ *is strictly convex, then* $\hat{\beta}_i = \hat{\beta}_j, \forall \lambda > 0$.

(b) If $P(\boldsymbol{\beta}) = |\boldsymbol{\beta}|$*, then* $\hat{\beta}_i \hat{\beta}_j \geq 0$ *and* $\hat{\boldsymbol{\beta}}'$ *is another maximizer of* $Q_n(\boldsymbol{\beta})$*, where*

$$\hat{\beta}'_k = \begin{cases} \hat{\beta}_k & \text{if } k \neq i \text{ and } k \neq j, \\ (\hat{\beta}_i + \hat{\beta}_j) \cdot (s) & \text{if } k = i, \\ (\hat{\beta}_i + \hat{\beta}_j) \cdot (1 - s) & \text{if } k = j. \end{cases}$$

for any $s \in [0, 1]$.

Proof. (a) Fix $\lambda > 0$. If $\hat{\beta}_i \neq \hat{\beta}_j$, let us consider $\hat{\boldsymbol{\beta}}'$ as follows:

$$\hat{\beta}'_k = \begin{cases} \hat{\beta}_k & \text{if } k \neq i \text{ and } k \neq j, \\ \dfrac{1}{2}(\hat{\beta}_i + \hat{\beta}_j) & \text{if } k = i \text{ or } k = j. \end{cases}$$

Because $\boldsymbol{x_i} = \boldsymbol{x_j}, i, j \in \{1, \ldots, p\}$, it is obvious that $\boldsymbol{x_l^T \hat{\beta}'} = \boldsymbol{x_l^T \hat{\beta}}$ for $l = 1, \ldots, n$; thus,

$$l(\hat{\boldsymbol{\beta}}') = \sum_{l=1}^{n} \{y_l(\boldsymbol{x_l}^T \hat{\boldsymbol{\beta}}') - b(\boldsymbol{x_l}^T \hat{\boldsymbol{\beta}}')\} = \sum_{l=1}^{n} \{y_l(\boldsymbol{x_l}^T \hat{\boldsymbol{\beta}}) - b(\boldsymbol{x_l}^T \hat{\boldsymbol{\beta}})\} = l(\hat{\boldsymbol{\beta}})$$

However, $P(\cdot)$ is strictly convex, so we have $P(\hat{\boldsymbol{\beta}}') < P(\hat{\boldsymbol{\beta}})$. Therefore $\hat{\boldsymbol{\beta}}$ cannot be the minimizer of (5), which is a contradiction. So we must have $\hat{\beta}_i = \hat{\beta}_j$.

(b) If $\hat{\beta}_i \hat{\beta}_j < 0$, Without loss of generality, assume $|\hat{\beta}_i| \geq |\hat{\beta}_j| > 0$, consider the same $\hat{\beta}'$ again,

$$|\hat{\boldsymbol{\beta}}'| = \sum_{k \neq i \; k \neq j} |\hat{\beta}_k| + \frac{1}{2}(|\hat{\beta}_i| - |\hat{\beta}_j|) + \frac{1}{2}(|\hat{\beta}_i| - |\hat{\beta}_j|) = \sum_{k=1}^{p} |\hat{\beta}_k| - 2|\hat{\beta}_j| = |\hat{\boldsymbol{\beta}}| - 2|\hat{\beta}_j| < |\hat{\boldsymbol{\beta}}|$$

So $\hat{\boldsymbol{\beta}}$ cannot be a Lasso solution. The rest can be directly verified by the definition of the Lasso, which is thus omitted.

□

This lemma applies to GLMs in both fixed p and diverging p_n cases, with or without sparsity assumptions.

4.2 CONSISTENCY WITHOUT SPARSITY ASSUMPTIONS

Fan and Li (2001) give a theorem regarding the existence and consistency of penalized likelihood estimators whose penalty part in (3), $\lambda P(\boldsymbol{\beta})$, can be written as $n \sum_{j=1}^{p} p_{\lambda_n}(|\beta_j|)$ in the fixed p case. Set $\boldsymbol{V}_i = (\boldsymbol{X_i}, Y_i), i = 1, \ldots, n$. Let $a_n = \max_{1 \leq j \leq s} p'_{\lambda_n}(|\beta_j^*|)$ and $b_n = \max_{1 \leq j \leq s} |p''_{\lambda_n}(|\beta_j^*|)|$.

Theorem 1 in Fan and Li (2001). *Let $\boldsymbol{V_1}, \ldots, \boldsymbol{V_n}$ be independent and identically distributed, each with a density $f(\boldsymbol{V}, \boldsymbol{\beta})$ (with respect to a measure μ) that satisfies conditions (C.1)-(C.3), and suppose that the penalty function $p_{\lambda_n}(\cdot)$ satisfies conditions (C.4) and (C.5). Then there exists a local maximizer $\hat{\boldsymbol{\beta}}_{\boldsymbol{n}}$ of $Q_n(\boldsymbol{\beta})$ such that $\|\hat{\boldsymbol{\beta}}_{\boldsymbol{n}} - \boldsymbol{\beta^*}\| = O_p(n^{-1/2})$.*

The regularity conditions are:

(C.1) $f(\boldsymbol{V}, \boldsymbol{\beta})$ has a common support and the model is identifiable. The first and second logarithmic derivatives of f satisfy the equations

$$E_{\boldsymbol{\beta}}\left\{\frac{\partial \log f(\boldsymbol{V},\boldsymbol{\beta})}{\partial \beta_j}\right\}=0 \quad \forall\, 1\le j\le p$$

and

$$\boldsymbol{I}_{jk}(\boldsymbol{\beta})=E_{\boldsymbol{\beta}}\left\{\frac{\partial \log f(\boldsymbol{V},\boldsymbol{\beta})}{\partial \beta_j}\frac{\partial \log f(\boldsymbol{V},\boldsymbol{\beta})}{\partial \beta_k}\right\}=E_{\boldsymbol{\beta}}\left\{-\frac{\partial^2 \log f(\boldsymbol{V},\boldsymbol{\beta})}{\partial \beta_j \partial \beta_k}\right\} \quad \forall\, 1\le j,k\le p.$$

(C.2) The Fisher information matrix

$$\boldsymbol{I}(\boldsymbol{\beta})=E_{\boldsymbol{\beta}}\left\{\left[\frac{\partial \log f(\boldsymbol{V},\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right]\left[\frac{\partial \log f(\boldsymbol{V},\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right]^T\right\}$$

is finite and positive definite at $\boldsymbol{\beta}=\boldsymbol{\beta}^*$.

(C.3) There exists a sufficiently large enough open set $\mathcal{O}$ that contains $\boldsymbol{\beta}^*$ such that for almost all $\boldsymbol{V}$ the density $f(\boldsymbol{V},\boldsymbol{\beta})$ admits all third derivatives $(\partial^3 f(\boldsymbol{V},\boldsymbol{\beta}))/(\partial\beta_j\partial\beta_k\partial\beta_l)$ for all $\boldsymbol{\beta}\in\mathcal{O}$. Furthermore, there exist functions M_{jkl} such that

$$\left|\frac{\partial^3 \log f(\boldsymbol{V},\boldsymbol{\beta})}{\partial \beta_j \partial \beta_k \partial \beta_l}\right| \le M_{jkl}(\boldsymbol{V}) \quad \forall\, \boldsymbol{\beta}\in\mathcal{O},$$

where $E_{\boldsymbol{\beta}}[M_{jkl}(\boldsymbol{V})]<\infty \quad \forall\, 1\le j,k,l\le p$.

(C.4) $a_n=O(n^{-1/2})$.

(C.5) $b_n\to 0$ as $n\to\infty$.

The regularity conditions (C.1)-(C.3) are similar to those that guarantee the usual asymptotics of MLEs.

Similarly, Fan and Peng (2004) give a theorem regarding the existence and consistency of penalized likelihood estimators whose penalty part in (3), $\lambda P(\boldsymbol{\beta})$, can be written as $n\sum_{j=1}^{p_n} p_{\lambda_n}(|\beta_j|)$ in the diverging p_n case. Set $\boldsymbol{V}_{ni}=(\boldsymbol{X_i},Y_{ni}), i=1,\ldots,n$. Let $a_n=\max_{1\le j\le s_n} p'_{\lambda_n}(|\beta^*_{nj}|)$ and $b_n=\max_{1\le j\le s_n}|p''_{\lambda_n}(|\beta^*_{nj}|)|$. Note that in the diverging p_n case here, both the true parameter vector β_n and the true nonzero parameter vector $\boldsymbol{\beta^*_{1n}}$ change as n grows, and the number of true parameter s_n may also change as n grows. There are no sparsity conditions (conditions on s_n) assumed here since we do not allow p_n to grow fast enough compared to n.

Theorem 1 in Fan and Peng (2004). *Let $\boldsymbol{V_{n1}}, \ldots, \boldsymbol{V_{nn}}$ be independent and identically distributed, each with a density $f_n(\boldsymbol{V_n}, \boldsymbol{\beta_n})$ (with respect to a measure μ) that satisfies conditions (C.1')-(C.3'), and suppose that the penalty function $p_{\lambda_n}(\cdot)$ satisfies conditions (C.4), (C.5), and (C.6'). If $p_n^4/n \to 0$ as $n \to \infty$, then there exists a local maximizer $\hat{\boldsymbol{\beta}}_{\boldsymbol{n}}$ of $Q_n(\boldsymbol{\beta})$ such that $\|\hat{\boldsymbol{\beta}}_{\boldsymbol{n}} - \boldsymbol{\beta}_{\boldsymbol{n}}^*\| = O_p((n/p_n)^{-1/2})$.*

The regularity conditions are:

(C.1') $f_n(\boldsymbol{V_n}, \boldsymbol{\beta_n})$ has a common support and the model is identifiable. The first and second logarithmic derivatives of f_n satisfy the equations

$$E_{\boldsymbol{\beta_n}}\left\{\frac{\partial \log f_n(\boldsymbol{V_n}, \boldsymbol{\beta_n})}{\partial \beta_{nj}}\right\} = 0 \quad \forall\, 1 \le j \le p_n$$

and

$$\begin{aligned}\boldsymbol{I_{njk}}(\boldsymbol{\beta_n}) &= E_{\boldsymbol{\beta_n}}\left\{\frac{\partial \log f_n(\boldsymbol{V_n}, \boldsymbol{\beta_n})}{\partial \beta_{nj}}\frac{\partial \log f_n(\boldsymbol{V_n}, \boldsymbol{\beta_n})}{\partial \beta_{nk}}\right\} \\ &= E_{\boldsymbol{\beta_n}}\left\{-\frac{\partial^2 \log f_n(\boldsymbol{V_n}, \boldsymbol{\beta_n})}{\partial \beta_{nj}\partial \beta_{nk}}\right\} \qquad \forall\, 1 \le j, k \le p_n.\end{aligned}$$

(C.2') The Fisher information matrix

$$\boldsymbol{I_n}(\boldsymbol{\beta_n}) = E_{\boldsymbol{\beta_n}}\left\{\left[\frac{\partial \log f_n(\boldsymbol{V_n}, \boldsymbol{\beta_n})}{\partial \boldsymbol{\beta_n}}\right]\left[\frac{\partial \log f_n(\boldsymbol{V_n}, \boldsymbol{\beta_n})}{\partial \boldsymbol{\beta_n}}\right]^T\right\}$$

satisfies conditions

$$0 < C_1 < \lambda_{min}\{\boldsymbol{I_n}(\boldsymbol{\beta_n})\} \le \lambda_{max}\{\boldsymbol{I_n}(\boldsymbol{\beta_n})\} < C_2 < \infty \quad \forall n,$$

and

$$E_{\boldsymbol{\beta_n}}\left\{\frac{\partial \log f_n(\boldsymbol{V_n}, \boldsymbol{\beta_n})}{\partial \beta_{nj}}\frac{\partial \log f_n(\boldsymbol{V_n}, \boldsymbol{\beta_n})}{\partial \beta_{nk}}\right\}^2 < C_3 < \infty \quad \forall\, 1 \le j, k \le p_n$$

$$E_{\boldsymbol{\beta_n}}\left\{\frac{\partial^2 \log f_n(\boldsymbol{V_n}, \boldsymbol{\beta_n})}{\partial \beta_{nj}\partial \beta_{nk}}\right\}^2 < C_4 < \infty \quad \forall\, 1 \le j, k \le p_n.$$

(C.3') There exists a sufficiently large enough open set $\mathcal{O}_n$ that contains $\boldsymbol{\beta}_{\boldsymbol{n}}^*$ such that for almost all $\boldsymbol{V_n}$ the density $f_n(\boldsymbol{V_n}, \boldsymbol{\beta_n})$ admits all third derivatives $(\partial^3 f_n(\boldsymbol{V_n}, \boldsymbol{\beta_n}))/(\partial\beta_{nj}\partial\beta_{nk}\partial\beta_{nl})$ for all $\boldsymbol{\beta_n} \in \mathcal{O}_n$. Furthermore, there exist functions M_{njkl} such that

$$\left|\frac{\partial^3 \log f_n(\boldsymbol{V_n}, \boldsymbol{\beta_n})}{\partial\beta_{nj}\partial\beta_{nk}\partial\beta_{nl}}\right| \leq M_{njkl}(\boldsymbol{V_n}) \quad \forall\, \boldsymbol{\beta_n} \in \mathcal{O}_n,$$

where $E_{\boldsymbol{\beta_n}}[M_{njkl}(\boldsymbol{V_n})]^2 < C_5 < \infty \quad \forall n \ \ \forall\, 1 \leq j, k, l \leq p_n$.

(C.6') There are constants C and D such that, when $\theta_1, \theta_2 > C\lambda_n$, $|p_{\lambda_n}^{''}(\theta_1) - p_{\lambda_n}^{''}(\theta_2)| \leq D|\theta_1 - \theta_2|$.

We consider the ridge, the Lasso, the elastic net, the adaptive Lasso, the adaptive elastic net and the SCAD estimators for GLMs. Based on Theorem 1 in Fan and Li (2001) and Theorem 1 in Fan and Peng (2004), we prove $\sqrt{n}$-consistency and $\sqrt{n/p_n}$-consistency of these penalized likelihood estimators for GLMs in the fixed p and diverging p_n cases, respectively. We give regularity conditions for these penalized likelihood estimators. Note again that in the diverging p_n case, both the true parameter vector β_n and the true nonzero parameter vector $\boldsymbol{\beta}_{\boldsymbol{1n}}^*$ change as n grows, and the number of true parameter s_n may also change as n grows. It still remains to show the consistency properties of these penalized likelihood estimators in the setup that p_n is of higher order than $o(n^{1/4})$.

Theorem 1. *Under the regularity conditions (A)-(C), there exists a local maximizer $\hat{\boldsymbol{\beta}}_{\boldsymbol{n}}$ of $Q_n(\boldsymbol{\beta})$ such that $\|\hat{\boldsymbol{\beta}}_{\boldsymbol{n}} - \boldsymbol{\beta}^*\| = O_p(n^{-1/2})$ in the fixed p case.*

The regularity conditions are:

(A) The Fisher information matrix

$$\boldsymbol{I}(\boldsymbol{\beta}) = E[b^{''}(\boldsymbol{x}^{\boldsymbol{T}}\boldsymbol{\beta})\boldsymbol{x}\boldsymbol{x}^{\boldsymbol{T}}]$$

is finite and positive definite at $\boldsymbol{\beta} = \boldsymbol{\beta}^*$.

(B) There exists a sufficiently large enough open set $\mathcal{O}$ that contains $\boldsymbol{\beta}^*$ such that $\forall \boldsymbol{\beta} \in \mathcal{O}$,

$$|b^{'''}(\boldsymbol{x}^{\boldsymbol{T}}\boldsymbol{\beta})| \leq M(\boldsymbol{x}) < \infty,$$

and

$$E[M(\boldsymbol{x})|x_j x_k x_l|] < \infty \quad \forall\, 1 \le j,k,l \le p.$$

(C) The penalty parameter(s) are of order $O(n^{1/2})$ for the ridge, the Lasso, the elastic net, the adaptive Lasso, and the adaptive elastic net. The penalty parameter $\lambda \to 0$ for the SCAD.

Proof. Note that

$$\begin{aligned}\log f(\boldsymbol{V},\boldsymbol{\beta}) &= \log[f_{y|\boldsymbol{x}}(y;\boldsymbol{\beta}) f_{\boldsymbol{x}}(\boldsymbol{x})] = \log f_{y|\boldsymbol{x}}(y;\boldsymbol{\beta}) + \log f_{\boldsymbol{x}}(\boldsymbol{x}) \\ &= y(\boldsymbol{x}^T\boldsymbol{\beta}) - b(\boldsymbol{x}^T\boldsymbol{\beta}) + C(\boldsymbol{x},y).\end{aligned}$$

Examine regularity conditions on likelihood functions (C.1)-(C.3) first,

(C.1) Obviously, $f(\boldsymbol{V},\boldsymbol{\beta})$ has a common support and the model is identifiable. And

$$\begin{aligned}E_{\boldsymbol{\beta}}\left\{\frac{\partial \log f(\boldsymbol{V},\boldsymbol{\beta})}{\partial \beta_j}\right\} &= E\{[y - b'(\boldsymbol{x}^T\boldsymbol{\beta})]x_j\} = E_{\boldsymbol{x}}\{E_{y|\boldsymbol{x}}[y - b'(\boldsymbol{x}^T\boldsymbol{\beta})]x_j\} \\ &= E_{\boldsymbol{x}}\{0 \cdot x_j\} = 0 \quad \forall\, 1 \le j \le p\end{aligned}$$

and

$$\begin{aligned}\boldsymbol{I_{jk}}(\boldsymbol{\beta}) &= E_{\boldsymbol{\beta}}\left\{\frac{\partial \log f(\boldsymbol{V},\boldsymbol{\beta})}{\partial \beta_j}\frac{\partial \log f(\boldsymbol{V},\boldsymbol{\beta})}{\partial \beta_k}\right\} = E\{[y - b'(\boldsymbol{x}^T\boldsymbol{\beta})]^2 x_j x_k\} \\ &= E_{\boldsymbol{x}}\{E_{y|\boldsymbol{x}}[y - b'(\boldsymbol{x}^T\boldsymbol{\beta})]^2 x_j x_k\} = E_{\boldsymbol{x}}\{b''(\boldsymbol{x}^T\boldsymbol{\beta}) x_j x_k\} \\ &= E_{\boldsymbol{\beta}}\left\{-\frac{\partial^2 \log f(\boldsymbol{V},\boldsymbol{\beta})}{\partial \beta_j \partial \beta_k}\right\} \quad \forall\, 1 \le j,k \le p.\end{aligned}$$

(C.2) From the proof of condition (C.1), we have

$$\boldsymbol{I}_{jk}(\boldsymbol{\beta}) = E_{\boldsymbol{x}}\{b''(\boldsymbol{x}^T\boldsymbol{\beta}) x_j x_k\},$$

thus

$$\boldsymbol{I}(\boldsymbol{\beta}) = E[b''(\boldsymbol{x^T\beta})\boldsymbol{x}\boldsymbol{x^T}].$$

This gives condition (A) in Theorem 1.

(C.3) It's clear that for all $\boldsymbol{V}$ the density $f(\boldsymbol{V}, \boldsymbol{\beta})$ admits all third derivatives $(\partial^3 f(\boldsymbol{V}, \boldsymbol{\beta}))/(\partial\beta_j\partial\beta_k\partial\beta_l)$. And that

$$\left|\frac{\partial^3 \log f(\boldsymbol{V}, \boldsymbol{\beta})}{\partial\beta_j\partial\beta_k\partial\beta_l}\right| = |-b'''(\boldsymbol{x}^T\boldsymbol{\beta})x_jx_kx_l| = |b'''(\boldsymbol{x}^T\boldsymbol{\beta})| \cdot |x_jx_kx_l|.$$

This gives condition (B) in Theorem 1.

We now investigate regularity conditions on penalty (C.4) and (C.5). They give condition (C) in Theorem 1. Note that we assume the true parameter vector $\boldsymbol{\beta}^*$ is fixed in the fixed p case.

1. Ridge: $\lambda P(\boldsymbol{\beta}) = \lambda\|\boldsymbol{\beta}\| = n\sum_{j=1}^p p_{\lambda_n}(|\beta_j|)$, then $p_{\lambda_n}(|\beta_j|) = \frac{\lambda}{n}|\beta_j|^2$, $p'_{\lambda_n}(|\beta_j|) = \frac{2\lambda}{n}|\beta_j|$, and $p''_{\lambda_n}(|\beta_j|) = \frac{2\lambda}{n}$. Thus,

$$a_n = \max_{1\le j\le s} p'_{\lambda_n}(|\beta_j^*|) = \frac{2\lambda}{n}\max_{1\le j\le s}|\beta_j^*| = O(n^{-1/2}) \implies \lambda = O(n^{1/2}),$$

$$b_n = \max_{1\le j\le s} |p''_{\lambda_n}(|\beta_j^*|)| = \frac{2\lambda}{n} \to 0 \implies \lambda = o(n).$$

Therefore, we need $\lambda = O(n^{1/2})$.

2. Lasso: $\lambda P(\boldsymbol{\beta}) = \lambda|\boldsymbol{\beta}| = n\sum_{j=1}^p p_{\lambda_n}(|\beta_j|)$, then $p_{\lambda_n}(|\beta_j|) = \frac{\lambda}{n}|\beta_j|$, $p'_{\lambda_n}(|\beta_j|) = \frac{\lambda}{n}$, and $p''_{\lambda_n}(|\beta_j|) = 0$. Thus,

$$a_n = \max_{1\le j\le s} p'_{\lambda_n}(|\beta_j^*|) = \frac{\lambda}{n} = O(n^{-1/2}) \implies \lambda = O(n^{1/2}),$$

$$b_n = 0.$$

Therefore, we need $\lambda = O(n^{1/2})$.

3. Elastic Net: $\lambda P_\alpha(\boldsymbol{\beta}) = \lambda(\alpha\|\boldsymbol{\beta}\|_2^2 + (1-\alpha)|\boldsymbol{\beta}|) = n\sum_{j=1}^p p_{\lambda_n}(|\beta_j|)$, then $p_{\lambda_n}(|\beta_j|) = \frac{\lambda}{n}(\alpha|\beta_j|^2 + (1-\alpha)|\beta_j|)$, $p'_{\lambda_n}(|\beta_j|) = \frac{\lambda}{n}(2\alpha|\beta_j| + 1 - \alpha)$, and $p''_{\lambda_n}(|\beta_j|) = \frac{2\lambda\alpha}{n}$. Thus,

$$a_n = \max_{1\le j\le s} p'_{\lambda_n}(|\beta_j^*|) = \lambda(\frac{2\alpha}{n}\max_{1\le j\le s}|\beta_j^*| + 1 - \alpha) = O(n^{-1/2}) \implies \lambda = O(n^{1/2}),$$

$$b_n = \max_{1\le j\le s} |p''_{\lambda_n}(|\beta_j^*|)| = \frac{2\lambda\alpha}{n} \to 0 \implies \lambda = o(n).$$

Therefore, we need $\lambda = O(n^{1/2})$.

4. Adaptive Lasso: $\lambda P(\boldsymbol{\beta}) = \sum_{j=1}^p \hat{w}_j|\beta_j| = n\sum_{j=1}^p p_{\lambda_n}(|\beta_j|)$, where $\hat{w}_j = |\hat{\beta}_j^{initial}|^{-\gamma}$ and $\hat{\boldsymbol{\beta}}^{initial}$ is a consistent estimator, then $p_{\lambda_n}(|\beta_j|) = \frac{\lambda}{n}\hat{w}_j|\beta_j|$, $p'_{\lambda_n}(|\beta_j|) = \frac{\lambda}{n}\hat{w}_j$, and $p''_{\lambda_n}(|\beta_j|) = 0$. Thus,

$$\begin{aligned} a_n &= \max_{1\le j\le s} p'_{\lambda_n}(|\beta_j^*|) = \frac{\lambda}{n}\max_{1\le j\le s}\hat{w}_j = \frac{\lambda}{n(\min_{1\le j\le s}|\hat{\beta}_j^{initial}|)^\gamma} \\ &\approx \frac{\lambda}{n(\min_{1\le j\le s}|\beta_j^*|)^\gamma} \quad \text{when } n \text{ is large} \\ &= O(n^{-1/2}) \implies \lambda = O(n^{1/2}), \end{aligned}$$

$$b_n = 0.$$

Therefore, we need $\lambda = O(n^{1/2})$.

5. Adaptive Elastic Net: $\lambda P(\boldsymbol{\beta}) = \lambda_1^*\sum_{j=1}^p \hat{w}_j|\beta_j| + \lambda_2^{enet}\|\boldsymbol{\beta}\|_2^2 = n\sum_{j=1}^p p_{\lambda_n}(|\beta_j|)$, where $\lambda_2^{enet} = \alpha^{enet}\lambda^{enet}$ and $\hat{w}_j = |\hat{\beta}_j^{enet}|^{-\gamma}$, then $p_{\lambda_n}(|\beta_j|) = \frac{\lambda_1^*}{n}\hat{w}_j|\beta_j| + \frac{\lambda_2^{enet}}{n}|\beta_j|^2$, $p'_{\lambda_n}(|\beta_j|) = \frac{\lambda_1^*}{n}\hat{w}_j + \frac{2\lambda_2^{enet}}{n}|\beta_j|$, and $p''_{\lambda_n}(|\beta_j|) = \frac{2\lambda_2^{enet}}{n}$. Thus,

$$\begin{aligned} a_n &= \max_{1\le j\le s} p'_{\lambda_n}(|\beta_j^*|) = \max_{1\le j\le s}\{\frac{\lambda_1^*}{n}\hat{w}_j + \frac{\lambda_2^{enet}}{n}|\beta_j^*|\} = \max_{1\le j\le s}\{\frac{\lambda_1^*}{n|\hat{\beta}_j^{enet}|^\gamma} + \frac{\lambda_2^{enet}}{n}|\beta_j^*|\} \\ &= \frac{\lambda_1^*}{n(\min_{1\le j\le s}|\hat{\beta}_j^{enet}|)^\gamma} + \frac{\lambda_2^{enet}}{n}\max_{1\le j\le s}|\beta_j^*| \\ &\approx \frac{\lambda_1^*}{n(\min_{1\le j\le s}|\beta_j^*|)^\gamma} + \frac{\lambda_2^{enet}}{n}\max_{1\le j\le s}|\beta_j^*| \quad \text{when } n \text{ is large and } \lambda^{enet} = O(n^{1/2}) \\ &= O(n^{-1/2}) \implies \lambda_1^* = O(n^{1/2}), \end{aligned}$$

$$b_n = \max_{1\le j\le s} |p''_{\lambda_n}(|\beta_j^*|)| = \frac{2\lambda_2^{enet}}{n} \to 0 \implies \lambda^{enet} = o(n).$$

Therefore, we need $\lambda_1^* = O(n^{1/2})$ and $\lambda^{enet} = O(n^{1/2})$.

6. SCAD: It is clear that if $\lambda \to 0$, then $a_n = 0$ and $b_n = 0$ when n is large enough. Therefore, conditions (C.4) and (C.5) are satisfied.

□

Theorem 2. *Under regularity conditions (A')-(C'), if $p_n^4/n \to 0$ as $n \to \infty$, then there exists a local maximizer $\hat{\boldsymbol{\beta}}_{\boldsymbol{n}}$ of $Q_n(\boldsymbol{\beta})$ such that $\|\hat{\boldsymbol{\beta}}_{\boldsymbol{n}} - \boldsymbol{\beta}_{\boldsymbol{n}}^*\| = O_p((n/p_n)^{-1/2})$ in the diverging p_n case.*

The regularity conditions are:

(A') The Fisher information matrix

$$\boldsymbol{I_n}(\boldsymbol{\beta_n}) = E[b''(\boldsymbol{x^T}\boldsymbol{\beta_n})\boldsymbol{x}\boldsymbol{x^T}]$$

satisfies condtions

$$0 < C_1 < \lambda_{min}\{\boldsymbol{I_n}(\boldsymbol{\beta_n})\} \le \lambda_{max}\{\boldsymbol{I_n}(\boldsymbol{\beta_n})\} < C_2 < \infty \quad \forall n$$

and

$$E\{[b^{(4)}(\boldsymbol{x^T}\boldsymbol{\beta_n}) + 3(b''(\boldsymbol{x^T}\boldsymbol{\beta_n}))^2]x_j^2x_k^2\} < C_3 < \infty \quad \forall\, 1 \le j,k \le p_n$$

$$E[b''(\boldsymbol{x^T}\boldsymbol{\beta_n})x_jx_k]^2 < C_4 < \infty \quad \forall\, 1 \le j,k \le p_n.$$

(B') There exists a sufficiently large enough open set $\mathcal{O}_n$ that contains $\boldsymbol{\beta}_{\boldsymbol{n}}^*$ such that $\forall \boldsymbol{\beta_n} \in \mathcal{O}_n$,

$$|b'''(\boldsymbol{x^T}\boldsymbol{\beta_n})| \le M_n(\boldsymbol{x}) < \infty$$

and

$$E[M_n(\boldsymbol{x})x_jx_kx_l]^2 < C_5 < \infty \quad \forall n \ \ \forall\, 1 \le j,k,l \le p_n.$$

(C') For the ridge, $\lambda = o(n)$ and $\lambda \max_{1\leq j\leq s_n} |\beta^*_{nj}| = O(n^{1/2})$.

For the Lasso, $\lambda = O(n^{1/2})$.

For the elastic net, $\lambda = O(n^{1/2})$ and $\lambda \max_{1\leq j\leq s_n} |\beta^*_{nj}| = O(n^{1/2})$.

For the adaptive Lasso, $\frac{\lambda}{(\min_{1\leq j\leq s_n} |\beta^*_{nj}|)^\gamma} = O(n^{1/2})$.

For the adaptive elastic net, $\lambda^{enet} = O(n^{1/2})$, $\lambda^{enet} \max_{1\leq j\leq s_n} |\beta^*_{nj}| = O(n^{1/2})$, and $\frac{\lambda_1^*}{(\min_{1\leq j\leq s_n} |\beta^*_{nj}|)^\gamma} = O(n^{1/2})$.

For the SCAD, $\lambda \to 0$ and $\min_{1\leq j\leq s_n} |\beta^*_{nj}|/\lambda \to \infty$.

Proof. As before,

$$\begin{aligned}\log f_n(\boldsymbol{V_n}, \boldsymbol{\beta_n}) &= \log[f_{y_n|\boldsymbol{x}}(y_n; \boldsymbol{\beta_n}) f_{\boldsymbol{x}}(\boldsymbol{x})] = \log f_{y_n|\boldsymbol{x}}(y_n; \boldsymbol{\beta_n}) + \log f_{\boldsymbol{x}}(\boldsymbol{x}) \\ &= y_n(\boldsymbol{x}^T\boldsymbol{\beta_n}) - b(\boldsymbol{x}^T\boldsymbol{\beta_n}) + C(\boldsymbol{x}, y_n).\end{aligned}$$

Next, examine regularity conditions on likelihood functions (C.1')-(C.3'),

(C.1') The same as in Theorem 1.

(C.2') From the proof of condition (C.1'), we have

$$\boldsymbol{I_{njk}}(\boldsymbol{\beta_n}) = E_{\boldsymbol{x}}\{b''(\boldsymbol{x}^T\boldsymbol{\beta_n})x_j x_k\},$$

thus

$$\boldsymbol{I_n}(\boldsymbol{\beta_n}) = E[b''(\boldsymbol{x^T}\boldsymbol{\beta_n})\boldsymbol{x}\boldsymbol{x^T}]$$

and

$$E_{\boldsymbol{\beta_n}}\left\{\frac{\partial \log f_n(\boldsymbol{V_n},\boldsymbol{\beta_n})}{\partial \beta_{nj}}\frac{\partial \log f_n(\boldsymbol{V_n},\boldsymbol{\beta_n})}{\partial \beta_{nk}}\right\}^2 = E\{[y-b'(\boldsymbol{x}^T\boldsymbol{\beta})]^2 x_j x_k\}^2$$
$$= E_{\boldsymbol{x}}\{E_{y|\boldsymbol{x}}[y-b'(\boldsymbol{x}^T\boldsymbol{\beta})]^4 x_j^2 x_k^2\} = E\{[b^{(4)}(\boldsymbol{x^T\beta_n})+3(b''(\boldsymbol{x^T\beta_n}))^2]x_j^2x_k^2\}$$

$$E_{\boldsymbol{\beta_n}}\left\{\frac{\partial^2 \log f_n(\boldsymbol{V_n},\boldsymbol{\beta_n})}{\partial \beta_{nj}\partial \beta_{nk}}\right\}^2 = E[b''(\boldsymbol{x}^T\boldsymbol{\beta})x_j x_k]^2$$

This gives condition (A') in Theorem 2.

(C.3') Similar as in Theorem 1. This gives condition (B') in Theorem 2.

We then investigate regularity conditions on penalty (C.4), (C.5), and (C.6'). It is obvious that (C.6') is satisfied for the ridge, the Lasso, the elastic net, the adaptive Lasso, and the adaptive elastic net. The proof of condition (C.4) and (C.5) in Theorem 1 gives condition (C') in Theorem 2 for these estimators. For the SCAD, if $\lambda \to 0$ and $\min_{1\le j\le s_n}|\beta^*_{nj}|/\lambda \to \infty$, then $a_n = 0$ and $b_n = 0$ when n is large enough. Therefore, conditions (C.4), (C.5), and (C.6') are satisfied.

□

Note that both Theorems 1 and 2 have limitations. In Theorem 1, we assume p to be fixed, which is not applicable to real world problems. In Theorem 2, we allow p_n to grow as n grows, but $p_n^4/n \to 0$ as $n \to \infty$ is also required. This means p_n can only be much smaller than n. In the next chapter, we will show a much stronger result where p_n can be much larger than n under the assumption that the true parameter vector is very sparse, in other words, s_n is much smaller than p_n.

4.3 NON-INFLUENTIAL VARIABLES AND THE RELAXED SENSE

Another big limitation of both Theorems 1 and 2 is that they are based on many regularity conditions that are hard to prove and easy to violate. For example, in Theorem 2, for the adaptive Lasso, the adaptive elastic net, and the SCAD, the minimum of nonzero parameters cannot converge to 0 too fast compared to λ. In other words, a sufficiently big smallest signal is needed to guarantee consistency. It is hard to prove and may be unrealistic.

Actually, it is usually not of interest to estimate those small parameters accurately. It will be more meaningful to treat those small parameters as exactly 0, so that the true model can be simplified. And some of those small parameters are so small that we may never be able to estimate them correctly using one specific estimator because of the consistency properties of the estimator. Treating small parameters as exactly 0 will also relax the regularity conditions, which will make penalized methods more applicable to real world problems.

Due to these considerations, we define non-influential parameters.

Definition 1. We call a variable ***non-influential*** if its coefficient grows slower than the rate of consistency of an estimator.

For example, in Theorem 2,

$$x_i \text{ is } \begin{cases} \text{non-influential} & \text{if } \beta_i^* = o((n/p_n)^{-1/2}) \\ \text{influential} & \text{if } \beta_i^* > o((n/p_n)^{-1/2}) \end{cases}$$

In other words, non-influential variables are those variables whose coefficients are comparably bigger and can make a difference. Here, I use the rate of consistency of an estimator as the threshold because it will be absolutely impossible for us to have accurate estimates of those variables, thus there is no point to try to make consistent estimates of those variables using the specific estimator. Different thresholds can be set to define non-influential variables based on the purposes of studies. And those thresholds should be higher than the rate of consistency of the estimator.

We can then define the relaxed sense based on the definition of non-influential variables.

Definition 2. We call a measure in the ***relaxed sense*** if the coefficients of non-influential variables are treated as zeros.

For example, in Theorem 2, let $\boldsymbol{\beta}_{\boldsymbol{r}}^*$ be the true parameter vector in the relaxed sense, whereas $\boldsymbol{\beta}^*$ is the true parameter vector in the usual sense. Then

$$\beta_{ir}^* = \begin{cases} 0 & \text{if } \beta_i^* = o((n/p_n)^{-1/2}) \\ \beta_i^* & \text{if } \beta_i^* > o((n/p_n)^{-1/2}) \end{cases}$$

The relaxed sense basically means we should treat small signals as irrelevant signals, because it is more worthwhile to ignore the small signals and focus on the bigger signals. Models will be much simpler in the relaxed sense and it is of our best interest to find the model which includes all

the bigger signals and excludes not only the irrelevant signals but also the small signals.

Regularity conditions and sparsity conditions can also be greatly loosened in the relaxed sense so that we can apply penalized likelihood methods to much more real world problems. Let us consider regularity condition (C') in Theorem 2. The relaxed sense would not change for the ridge, the Lasso, and the elastic net. It would change for the adaptive Lasso, the adaptive elastic net, and the SCAD, as

$$\min_{1\le j\le s_n} |\beta^*_{nj}|(\text{relaxed}) = \min_{1\le j\le s_n} |\beta^*_{nrj}| = \min_{1\le j\le s_n:\beta^*_{nj}\neq o((n/p_n)^{-1/2})} |\beta^*_{nj}|.$$

Then, in the relaxed sense, condition (C') in Theorem 2 is loosened to

For the adaptive Lasso,

$$\frac{\lambda}{(\min_{1\le j\le s_n:\beta^*_{nj}\neq o((n/p_n)^{-1/2})} |\beta^*_{nj}|)^\gamma} = O(n^{1/2}).$$

For the adaptive elastic net, $\lambda^{enet} = O(n^{1/2})$, $\lambda^{enet} \max_{1\le j\le s_n} |\beta^*_{nj}| = O(n^{1/2})$, and

$$\frac{\lambda_1^*}{(\min_{1\le j\le s_n:\beta^*_{nj}\neq o((n/p_n)^{-1/2})} |\beta^*_{nj}|)^\gamma} = O(n^{1/2}).$$

For the SCAD, $\lambda \to 0$ and

$$\min_{1\le j\le s_n:\beta^*_{nj}\neq o((n/p_n)^{-1/2})} |\beta^*_{nj}|/\lambda \to \infty.$$

4.4 A SIMULATION STUDY

We perform a simulation study to compare the performance of the ridge, the Lasso, the elastic net, the adaptive Lasso, the adaptive elastic net, and the SCAD estimators in the usual and relaxed senses. We consider the penalized likelihood estimators for logistic models,

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \{-y_i(\boldsymbol{x_i}^T\boldsymbol{\beta}) + \log(1 + e^{\boldsymbol{x_i}^T\boldsymbol{\beta}})\} + \lambda P(\boldsymbol{\beta}) \tag{11}$$

where $y_i, i = 1, \ldots, n$ is 1 or 0.

4.4.1 Methods

The MLE can be used for the estimated weights for the adaptive Lasso in the $n > p$ case. However, it is nontrivial to find an appropriate estimate when p is comparable to or larger that n. A practical solution is to use the ridge estimator, as it is more stable than the MLE for collinearity problem. Thus the adaptive Lasso can be well defined. In all simulations, we compute the adaptive weights using ridge estimates.

Followed the suggestion of Fan and Li (2001) that based on a Bayesian point of view and simulation studies, $a = 3.7$ is used in the SCAD estimator.

4.4.2 Implementation

We use the coordinate descent algorithm (Friedman et al., 2010) to compute the ridge, the Lasso, the elastic net and the adaptive Lasso estimators. We implement the LQA algorithm of Fan and Li (2001) to compute the adaptive elastic net and the SCAD estimators.

For each competitor, we select its tuning parameter(s) by tenfold cross-validation based on the binomial deviance. We use the binomial deviance that is twice the negative log-likelihood rather than misclassification error, since the deviance is smoother. Cross-validation is used to select the mixing parameter α as well in the simulations, although it is often viewed as a higher-level parameter and chosen on more subjective grounds.

For each simulation, we simulate 100 datasets consisting of n observations from a logistic model. The predictors $\boldsymbol{x_i}, i = 1, \ldots, n$ are iid normal vectors. We set the pairwise correlation between x_{j_1} and x_{j_2} to be $\text{cor}(x_{j_1}, x_{j_2}) = \rho^{|j_1 - j_2|}$ with $\rho = .5$. The mean of each predictor is set to be 0 and the variance is set to be 1. We do not include the intercept in the candidate models because the expectation of the simulated responses is 0.5, which means we have balanced 1s and 0s on average.

4.4.3 Measures

Denote the number of true positive parameters as $s+$, the number of true negative parameters as $s-$, and the number of true zero parameters as p_0 (which is equal to $p - s$).

For each estimator $\hat{\boldsymbol{\beta}}$, its parameter estimation accuracy is measured by the mean L_2 loss from the true parameter $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2$ (ML2). The variable selection performance is gauged by the mean of the number of true positive variables that are correctly estimated as positive $\sum_{i:\beta_i^*>0} I(\hat{\beta}_i > 0)$ (MC+), the mean of the number of true negative variables that are correctly estimated as negative $\sum_{i:\beta_i^*<0} I(\hat{\beta}_i < 0)$ (MC-), the mean of the number of true zero variables that are correctly estimated as zero $\sum_{i:\beta_i^*=0} I(\hat{\beta}_i = 0)$ (MC0), and the percent time of getting all the signs correctly $I(\mathrm{sgn}(\hat{\boldsymbol{\beta}}) = \mathrm{sgn}(\boldsymbol{\beta}^*))$ (%PS).

We also report their counterparts in the relaxed sense. Here we call a measure in the relaxed sense if the small nonzero true parameters that are of order $o(n^{-1/2})$ are treated as exactly 0.

4.4.4 Models

We consider the following four models for $n = 40$ and 80. The dimensionality p is fixed.

Model 1. In this model, we let $\boldsymbol{\beta}^* = (3, 1.5, 0, 0, 2, 0, 0, 0)^T$. Thus, $p = 8, s = 3, s+ = 3, s- = 0$, and $p_0 = 5$. It is sparse and stays the same in the relaxed sense.

Model 2. In this model, we let $\boldsymbol{\beta}^* = (3, 1.5, 1/80, 1/80, 2, 1/80, 1/80, 1/80)^T$. Thus, $p = 8, s = 8, s+ = 8, s- = 0$, and $p_0 = 0$. However, in the relaxed sense, $\boldsymbol{\beta}^* = (3, 1.5, 0, 0, 2, 0, 0, 0)^T$ as in Model 1, and it is sparse in the relaxed sense with $p = 8, s = 3, s+ = 3, s- = 0$, and $p_0 = 5$

Model 3. In this model, we let $\boldsymbol{\beta}^* = (3, -1.5, 0, 0, 2, 0, 0, 0)^T$. Thus, $p = 8, s = 3, s+ = 2, s- = 1$, and $p_0 = 5$. It is sparse and stays the same in the relaxed sense.

Model 4. In this model, we let $\boldsymbol{\beta}^* = (3, -1.5, -1/80, 1/80, 2, 1/80, 1/80, -1/80)^T$. Thus, $p = 8, s = 8, s+ = 5, s- = 3$, and $p_0 = 0$. However, in the relaxed sense, $\boldsymbol{\beta}^* = (3, -1.5, 0, 0, 2, 0, 0, 0)^T$ as in Model 3, and it is sparse in the relaxed sense with $p = 8, s = 3, s+ = 2, s- = 1$, and $p_0 = 5$.

4.4.5 Results

The simulation results are presented in Tables 1–6.

Table 1. Simulation Results for Model 1 based on 100 realizations (Usual)

Method	ML2	MC+	MC-	MC0	%PS
$n = 40$					
True	**0**	**3**	**0**	**5**	**1**
Ridge	2.2766	3	0	0	0
Lasso	2.7339	2.9	0	2.59	0.09
Elastic Net	2.7880	2.96	0	1.58	0.01
Adaptive Lasso	3.8861	2.68	0	3.82	0.24
Adaptive Elastic Net	2.5882	2.96	0	1.58	0.01
SCAD	2.7296	2.98	0	0	0
$n = 80$					
True	**0**	**3**	**0**	**5**	**1**
Ridge	1.6312	3	0	0	0
Lasso	1.5412	2.99	0	2.29	0.05
Elastic Net	1.6130	3	0	1.57	0.02
Adaptive Lasso	1.3749	2.91	0	4.15	0.43
Adaptive Elastic Net	1.5364	3	0	1.57	0.02
SCAD	1.3996	3	0	0	0

Table 2. Simulation Results for Model 2 based on 100 realizations (Usual)

Method	ML2	MC+	MC-	MC0	%PS
$n = 40$					
True	**0**	**8**	**0**	**0**	**1**
Ridge	2.3025	5.92	0	0	0.03
Lasso	2.2789	4.46	0	0	0
Elastic Net	2.5049	5.17	0	0	0
Adaptive Lasso	3.8931	3.51	0	0	0
Adaptive Elastic Net	2.1307	5.09	0	0	0.02
SCAD	9.4421	6.26	0	0	0.2
$n = 80$					
True	**0**	**8**	**0**	**0**	**1**
Ridge	1.5980	5.83	0	0	0.02
Lasso	1.4555	4.73	0	0	0.01
Elastic Net	1.5411	5.14	0	0	0.02
Adaptive Lasso	1.5080	3.58	0	0	0
Adaptive Elastic Net	1.5001	5.12	0	0	0.02
SCAD	1.6907	5.61	0	0	0.09

Table 3. Simulation Results for Model 2 based on 100 realizations (**Relaxed**)

Method	ML2	MC+	MC-	MC0	%PS
$n = 40$					
True	**0**	**3**	**0**	**5**	**1**
Ridge	2.3079	2.97	0	0	0
Lasso	2.2830	2.9	0	2.71	0.05
Elastic Net	2.5102	2.95	0	1.54	0.04
Adaptive Lasso	3.8960	2.72	0	3.85	0.27
Adaptive Elastic Net	2.1346	2.95	0	1.54	0.04
SCAD	9.4432	2.97	0	0	0
$n = 80$					
True	**0**	**3**	**0**	**5**	**1**
Ridge	1.5999	2.99	0	0	0
Lasso	1.4574	2.99	0	2.34	0.1
Elastic Net	1.5431	2.99	0	1.5	0.04
Adaptive Lasso	1.5091	2.96	0	4.08	0.44
Adaptive Elastic Net	1.5020	2.99	0	1.5	0.04
SCAD	1.6911	2.99	0	0	0

Table 4. Simulation Results for Model 3 based on 100 realizations (Usual)

Method	ML2	MC+	MC-	MC0	%PS
$n = 40$					
True	**0**	**2**	**1**	**5**	**1**
Ridge	2.1927	2	0.96	0	0
Lasso	2.1279	1.99	0.8	2.49	0.01
Elastic Net	2.3420	2	0.88	1.35	0.01
Adaptive Lasso	2.9976	2	0.72	3.77	0.2
Adaptive Elastic Net	2.0479	2	0.88	1.35	0.01
SCAD	2.6584	2	0.93	0	0
$n = 80$					
True	**0**	**2**	**1**	**5**	**1**
Ridge	1.5837	2	1	0	0
Lasso	1.4894	2	0.97	2.12	0.05
Elastic Net	1.5519	2	0.98	1.11	0.01
Adaptive Lasso	1.4359	2	0.89	3.83	0.36
Adaptive Elastic Net	1.5165	2	0.98	1.11	0.01
SCAD	1.8488	2	1	0	0

Table 5. Simulation Results for Model 4 based on 100 realizations (Usual)

Method	ML2	MC+	MC-	MC0	%PS
$n = 40$					
True	**0**	**5**	**3**	**0**	**1**
Ridge	2.2045	3.9	2.14	0	0.08
Lasso	2.0780	2.89	1.33	0	0
Elastic Net	2.1499	3.3	1.56	0	0.01
Adaptive Lasso	2.1794	2.4	1	0	0
Adaptive Elastic Net	2.0337	3.27	1.57	0	0
SCAD	2.4386	3.65	2.28	0	0.08
$n = 80$					
True	**0**	**5**	**3**	**0**	**1**
Ridge	1.5923	3.67	2.12	0	0.08
Lasso	1.5331	2.98	1.6	0	0
Elastic Net	1.6167	3.29	1.89	0	0.03
Adaptive Lasso	1.3262	2.41	1.19	0	0
Adaptive Elastic Net	1.4954	3.24	1.9	0	0.02
SCAD	1.7625	3.32	2.26	0	0.03

Table 6. Simulation Results for Model 4 based on 100 realizations (**Relaxed**)

Method	ML2	MC+	MC-	MC0	%PS
$n = 40$					
True	**0**	**2**	**1**	**5**	**1**
Ridge	2.2078	2	0.97	0	0
Lasso	2.0795	1.99	0.81	2.79	0.06
Elastic Net	2.1512	2	0.86	1.71	0.03
Adaptive Lasso	2.1802	1.99	0.71	3.95	0.25
Adaptive Elastic Net	2.0354	2	0.86	1.71	0.03
SCAD	2.4388	2	0.98	0	0
$n = 80$					
True	**0**	**2**	**1**	**5**	**1**
Ridge	1.5934	2	0.99	0	0
Lasso	1.5339	2	0.95	2.32	0.02
Elastic Net	1.6175	2	0.98	1.22	0.01
Adaptive Lasso	1.3268	2	0.92	3.88	0.34
Adaptive Elastic Net	1.4962	2	0.98	1.22	0.01
SCAD	1.7633	2	1	0	0

4.4.6 Discussion

Several interesting observations can be made:

1. The performance of the estimators for the dense models in the relaxed sense is comparable to their performance for the sparse models, regardless of whether there are negative parameters involved. The estimators do worse in identifying the sign of the parameters for the dense model in the usual sense.

2. The parameter estimation performance of the estimators is better when the sample size gets larger for the sparse models and the dense models (in both the usual sense and the relaxed sense). It appears that the estimators are consistent for parameter estimation as Theorem 1 suggests. The variable selection performance of most of the estimators improves slowly as the sample size gets larger.

3. For the sparse models and the dense models in the relaxed sense, the adaptive Lasso performs the best, especially when the sample size is not very small ($n = 80$). It does much better than the others in identifying the true zero parameters and finding the right signs for all the parameters. The Lasso also outperforms the elastic net and the adaptive elastic net. It is perhaps because the correlation is only moderate. We suspect that the adaptive Lasso and the Lasso may not perform as well when the correlation gets higher (say, $\rho = .7$).

4. The SCAD estimator never identifies the true zero parameters. It does a poor job in selecting variables.

5.0 CONSISTENCY WITH SPARSITY ASSUMPTIONS

5.1 "LARGE P, SMALL N" PROBLEMS

In the last chapter, we either require a fixed p or only allow p_n to grow much slower than n, though, in many applications, it is more common that p_n grows much faster than n. Hence, we should give more attention to "Large p, Small n" problems. Here, we show a much stronger result where p_n can be much larger than n under the assumption that the true parameter vector is very sparse, in other words, s_n is much smaller than p_n. Note that again we assume both the true parameter vector β_n and the true nonzero parameter vector $\boldsymbol{\beta}_{\mathbf{1}\boldsymbol{n}}^{*}$ change as n grows, and here we assume the number of true parameters s_n grows as n grows.

Fan and Lv (2011) prove that for nonconcave penalties, like the Lasso and the SCAD penalties, the penalized estimators are $sqrtn/s_n$ consistent and have Oracle Property under some regularity conditions.

Theorem 3 & 4 in Fan and Lv (2011). *When* $\log p_n = O(n^a)$ *for some* $a \in (0,1)$ *and* $s_n = o(n)$*, under some regularity conditions, for a nonconcave penalty, there exists a local maximizer* $\hat{\boldsymbol{\beta}}_{\boldsymbol{n}}$ *of* $Q_n(\boldsymbol{\beta})$ *such that* $\|\hat{\boldsymbol{\beta}}_{\boldsymbol{n}} - \boldsymbol{\beta}_{\boldsymbol{n}}^{*}\| = O_p((n/s_n)^{-1/2})$*; and if* $s_n = o(n^{1/3})$*, the maximizer* $\hat{\boldsymbol{\beta}}_{\boldsymbol{n}}$ *has the oracle property.*

Note that p_n is allowed to grow at an exponential rate, but s_n is required to grow much slower than n. We therefore assume the true parameter vector to be very sparse. By checking the regularity conditions, the relaxed sense does not seem to be able to relax the regularity conditions, but it will greatly relax the sparsity conditions.

In the usual sense, the sparsity s_n is defined as

$$s_n = \sum_{1 \le i \le p_n} I(\beta_i^* \neq 0) \tag{12}$$

In the relaxed sense,

$$\beta_{ri}^* = \begin{cases} 0 & \text{if } \beta_i^* = o((n/s_n)^{-1/2}) \\ \beta_i^* & \text{if } \beta_i^* > o((n/s_n)^{-1/2}) \end{cases}$$

and the sparsity s_n becomes

$$s_{rn} = \sum_{1 \le i \le p_n} I(\beta_{ri}^* \neq 0) \tag{13}$$

Therefore, the sparsity condition can be greatly loosened when p_n is large, especially when there are a lot of small effects.

5.2 A SIMULATION STUDY

5.2.1 Setup

The general simulation setup is the same as that in Chapter 4, but we consider "Big p, Small n" models. Since including negative parameters does not seem to make a difference in the simulation study in the last chapter, we do not include negative parameters in the simulation studies in this chapter.

Take $a = .5$ in $\log p_n = O(n^a)$.

Model 5. In this model, we let $n = 40$, $p = 600$, and

$$\boldsymbol{\beta^*} = (\mathbf{3}, \mathbf{1.5}, \frac{3}{n}, \frac{3}{n}, \mathbf{2}, \frac{3}{n}, \cdots, \frac{3}{n})^T.$$

Thus, $p = 600, s = 600$, and $p_0 = 0$. However, in the relaxed sense,

$$\boldsymbol{\beta^*} = (\mathbf{3}, \mathbf{1.5}, 0, 0, \mathbf{2}, 0, \cdots, 0)^T.$$

It is sparse in the relaxed sense with $p = 600, s = 3$, and $p_0 = 597$.

Model 6. In this model, we let $n = 80$, $p = 8000$, and

$$\boldsymbol{\beta}^* = (\mathbf{3}, \mathbf{1.5}, \frac{5}{n}, \frac{5}{n}, \mathbf{2}, \frac{5}{n}, \frac{5}{n}, \frac{5}{n}, \mathbf{2}, \mathbf{2}, \frac{5}{n}, \cdots, \frac{5}{n})^T.$$

Thus, $p = 8000, s = 8000$, and $p_0 = 0$. However, in the relaxed sense,

$$\boldsymbol{\beta}^* = (\mathbf{3}, \mathbf{1.5}, 0, 0, \mathbf{2}, 0, 0, 0, \mathbf{2}, \mathbf{2}, 0, \cdots, 0)^T.$$

It is sparse in the relaxed sense with $p = 8000, s = 5$, and $p_0 = 7995$.

For each simulation, we simulate 100 datasets consisting of n observations from a logistic model. The predictors $\boldsymbol{x_i}, i = 1, \ldots, n$ are iid normal vectors. We set the pairwise correlation between x_{j_1} and x_{j_2} to be $\text{cor}(x_{j_1}, x_{j_2}) = \rho^{|j_1 - j_2|}$ with $\rho = .3, .5$, and $.7$. We do not include negative parameters because the simulations in last chapter show that the performance is not influenced by the sign of the parameters. Therefore, we do not report measure MC-.

5.2.2 Results

The simulation results are presented in Tables 7 - 12.

Table 7. Simulation Results for Models 5 & 6 based on 100 realizations (Usual)

Method	ML2	MC+	MC0	%PS
$\rho = .5$				
$n = 40$				
True	**0**	**600**	**0**	**1**
Ridge	2.5251	6.21	0	0
Lasso	2.9009	4.73	0	0
Elastic Net	2.6246	5.43	0	0
Adaptive Lasso	3.6366	3.7	0	0
Adaptive Elastic Net	2.5771	5.43	0	0
SCAD	10.04067	6.478	0	0.01
$n = 80$				
True	**0**	**8000**	**0**	**1**
Ridge	2.6201	11.05	0	0.08
Lasso	2.3573	8.37	0	0
Elastic Net	2.5404	9.35	0	0.03
Adaptive Lasso	2.5665	6.4	0	0
Adaptive Elastic Net	2.4680	9.31	0	0.06
SCAD	2.6205	10.98	0	0.3

Table 8. Simulation Results for Models 5 & 6 based on 100 realizations (**Relaxed**)

Method	ML2	MC+	MC0	%PS
$\rho = .5$				
$n = 40$				
True	**0**	**3**	**597**	**1**
Ridge	2.5440	2.98	0	0
Lasso	2.9156	2.86	592.61	0
Elastic Net	2.6416	2.95	591.45	0
Adaptive Lasso	3.6449	2.73	592.78	0.17
Adaptive Elastic Net	2.5920	2.94	591.45	0
SCAD	10.0403	2.96	0	0
$n = 80$				
True	**0**	**5**	**7995**	**1**
Ridge	2.6327	4.99	0	0
Lasso	2.3646	4.99	7991.93	0.06
Elastic Net	2.5507	4.99	7990.29	0.01
Adaptive Lasso	2.5692	4.9	7994.77	0.29
Adaptive Elastic Net	2.4753	4.99	7990.29	0.01
SCAD	2.6186	4.99	0	0

Table 9. Simulation Results for Models 5 & 6 based on 100 realizations (Usual)

Method	ML2	MC+	MC0	%PS
$\rho = .3$				
$n = 40$				
True	**0**	**600**	**0**	**1**
Ridge	2.2875	5.9	0	0.02
Lasso	2.4273	4.62	0	0
Elastic Net	2.4683	5.18	0	0
Adaptive Lasso	2.9805	3.7	0	0
Adaptive Elastic Net	2.3472	5.08	0	0
SCAD	2.4235	5.9	0	0.01
$n = 80$				
True	**0**	**8000**	**0**	**1**
Ridge	2.4081	10.59	0	0.07
Lasso	2.5451	8.37	0	0
Elastic Net	2.4590	9.16	0	0.03
Adaptive Lasso	3.3699	6.38	0	0
Adaptive Elastic Net	2.2822	9.28	0	0.02
SCAD	2.6365	10.91	0	0.15

Table 10. Simulation Results for Models 5 & 6 based on 100 realizations (**Relaxed**)

Method	ML2	MC+	MC0	%PS
$\rho = .3$				
$n = 40$				
True	**0**	**3**	**597**	**1**
Ridge	2.2958	3	0	0
Lasso	2.4304	2.89	592.31	0
Elastic Net	2.4729	2.97	591.41	0
Adaptive Lasso	2.9810	2.84	593.6	0.18
Adaptive Elastic Net	2.3493	2.97	591.41	0
SCAD	2.41941	2.99	0	0
$n = 80$				
True	**0**	**5**	**7995**	**1**
Ridge	2.4151	5	0	0
Lasso	2.5497	4.99	7991.67	0.09
Elastic Net	2.4647	4.99	7989.9	0.04
Adaptive Lasso	3.3688	4.93	7994.88	0.25
Adaptive Elastic Net	2.2849	4.99	7989.9	0.04
SCAD	2.6316	5	0	0

Table 11. Simulation Results for Models 5 & 6 based on 100 realizations (Usual)

Method	ML2	MC+	MC0	%PS
$\rho = .7$				
$n = 40$				
True	**0**	**600**	**0**	**1**
Ridge	2.4894	6.09	0	0
Lasso	2.6042	4.64	0	0
Elastic Net	2.5878	5.16	0	0
Adaptive Lasso	5.0850	3.51	0	0
Adaptive Elastic Net	2.4859	5.21	0	0
SCAD	2.9931	6.61	0	0.13
$n = 80$				
True	**0**	**8000**	**0**	**1**
Ridge	2.8495	11.33	0	0.07
Lasso	3.1141	8.24	0	0
Elastic Net	2.9245	9.68	0	0.02
Adaptive Lasso	3.5104	6.27	0	0
Adaptive Elastic Net	3.0142	9.73	0	0.04
SCAD	3.6406	11.89	0	0.35

Table 12. Simulation Results for Models 5 & 6 based on 100 realizations (**Relaxed**)

Method	ML2	MC+	MC0	%PS
$\rho = .7$				
$n = 40$				
True	**0**	**3**	**597**	**1**
Ridge	2.5096	2.97	0	0
Lasso	2.6217	2.81	592.71	0
Elastic Net	2.6071	2.89	591.91	0
Adaptive Lasso	5.0979	2.56	593.79	0.14
Adaptive Elastic Net	2.5029	2.9	591.91	0
SCAD	2.9989	2.95	0	0
$n = 80$				
True	**0**	**5**	**7995**	**1**
Ridge	2.8737	4.99	0	0
Lasso	3.1325	4.88	7992.67	0.04
Elastic Net	2.9459	4.94	7990.15	0.03
Adaptive Lasso	3.5230	4.58	7994.83	0.17
Adaptive Elastic Net	3.0327	4.94	7990.15	0.03
SCAD	3.6433	4.94	0	0

5.2.3 Discussion

Several interesting observations can be made:

1. The performance of the estimators in the relaxed sense is much better than in the usual sense, except the performance of the SCAD with regard to %PS. The SCAD tends to pick out variables with small nonzero parameters.

2. The parameter estimation performance of the estimators is better when the sample size gets larger in both the usual sense and the relaxed sense. It appears that the estimators are consistent for parameter estimation. The variable selection performance of most of the estimators does not improve much when the sample size gets larger.

3. The adaptive Lasso performs the best in the relaxed sense, especially when the sample size

is not very small ($n = 80$). It does much better than the others in identifying the true zero parameters and finding the right signs for all the parameters. The Lasso also outperforms the elastic net and the adaptive elastic net.

4. The Elastic Net and Adaptive Elastic Net do well with respect to the measures ML2, MC+, and MC0 in the relaxed sense, although they are not nonconcave penalties.

5. The variable selection performance of the penalized likelihood estimators seem to get worse when correlations get stronger. Though, the differences are small, and may be due to pure chance.

6. The Lasso, Elastic Net, Adaptive Lasso, and Adaptive Elastic Net do well with respect to the measures ML2, MC+, and MC0 in the relaxed sense, but have poor performances in regard to %PS. In other words, they have parameter estimation consistency, and identify the big parameters on average, but if we look at each set of parameter estimates, it often misses some big variables with parameters or includes some variables with small parameters in the model.

5.3 A SIMULATION STUDY WITH KNOWN X STRUCTURE

5.3.1 Gene expression data

We consider a gene expression data set that consists of 72 leukemia patients and 7192 variables (Golub et al., 1999). 6817 of them are gene expression levels, the others are RNA or protein levels. There are 47 patients who have acute lymphoblastic leukemia (ALL) and 25 patients who have acute myeloid leukemia (AML). The goal is to find the important variables that distinguish AML and ALL.

Below is a heat map of part of the data, with 70 gene expression levels of 38 patients. Each row represents a gene, and each column represents a patient.

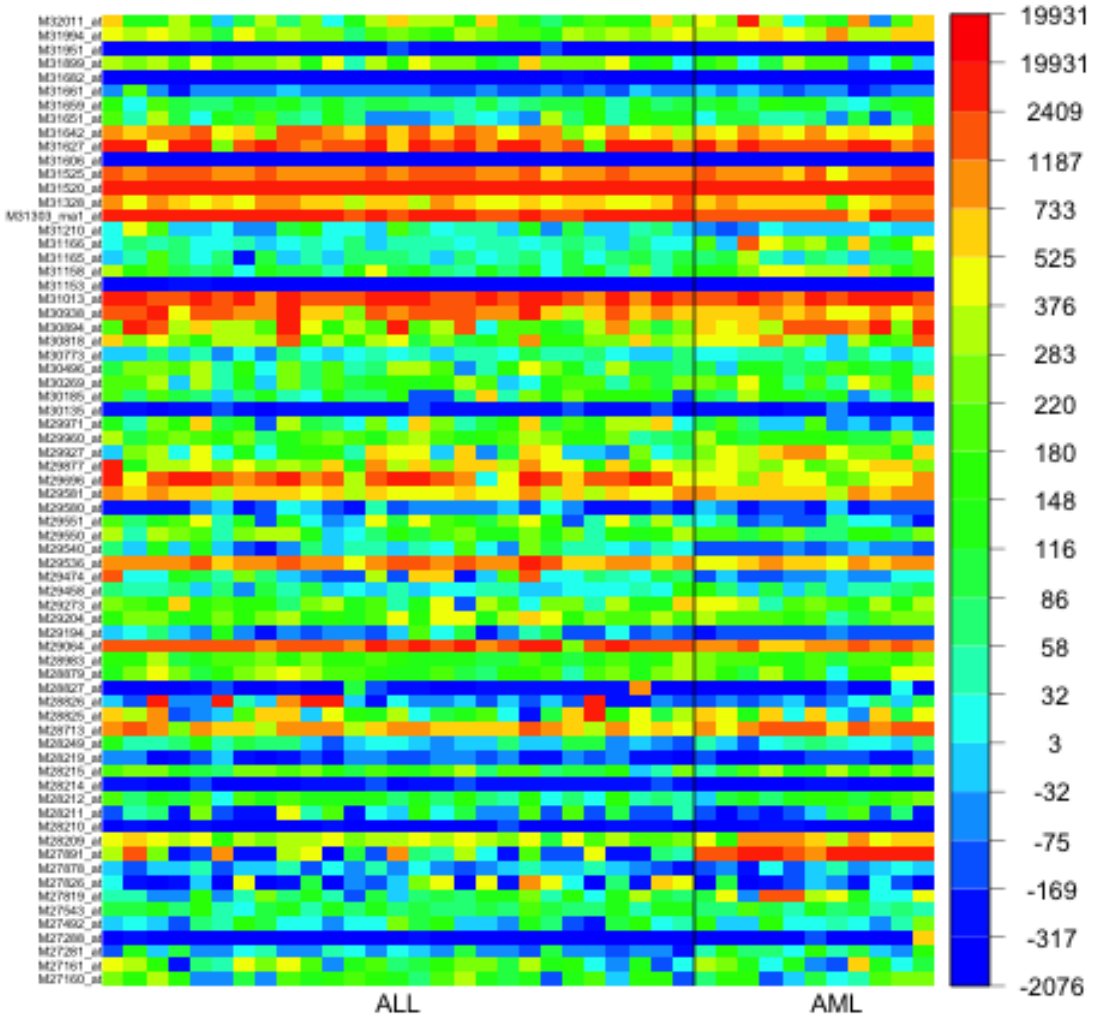

Figure 1: Partial Heatmap of 70 gene expression levels of 38 patients. ($n = 72, p = 7,192$)

Most genes express similarly in AML patients and ALL patients. But it appears that genes M27819, M27878, and M27891 express more in AML patients than ALL patients.

5.3.2 Setup

In the previous section, we try difference correlation structures in the simulation. Here, we use the estimated correlation structure of this data set and then simulate 100 marginal standard normal variables based on the pre-specified correlation structure.

Model 7. In this model, we let $n = 72$, $p = 7192$, and

$$\boldsymbol{\beta}^* = (\mathbf{3}, \mathbf{1.5}, \frac{5}{n}, \frac{5}{n}, \mathbf{2}, \frac{5}{n}, \frac{5}{n}, \frac{5}{n}, \mathbf{2}, \mathbf{2}, \frac{5}{n}, \cdots, \frac{5}{n})^T.$$

Thus, $p = 7192, s = 7192$, and $p_0 = 0$. However, in the relaxed sense,

$$\boldsymbol{\beta}^* = (\mathbf{3}, \mathbf{1.5}, 0, 0, \mathbf{2}, 0, 0, 0, \mathbf{2}, \mathbf{2}, 0, \cdots, 0)^T.$$

It is sparse in the relaxed sense with $p = 7192, s = 5$, and $p_0 = 7187$.

We only include big and small effects in our previous simulations. Here, we also try a model that has medium effects along with big and small eftects.

Model 8. In this model, we let $n = 72$, $p = 7192$, and

$$\boldsymbol{\beta}^* = (\mathbf{3}, \mathbf{1.5}, \frac{5}{n}, \mathbf{.5}, \mathbf{2}, \frac{5}{n}, \frac{5}{n}, \frac{5}{n}, \mathbf{2}, \mathbf{2}, \frac{5}{n}, \frac{5}{n}, \frac{5}{n}, \frac{5}{n}, \mathbf{.5}, \frac{5}{n}, \cdots, \frac{5}{n})^T.$$

Thus, $p = 7192, s = 7192$, and $p_0 = 0$. However, in the relaxed sense,

$$\boldsymbol{\beta}^* = (\mathbf{3}, \mathbf{1.5}, 0, \mathbf{.5}, \mathbf{2}, 0, 0, 0, \mathbf{2}, \mathbf{2}, 0, 0, 0, 0, \mathbf{.5}, 0, \cdots, 0)^T.$$

It is sparse in the relaxed sense with $p = 7192, s = 7$, and $p_0 = 7185$.

5.3.3 Results

The simulation results are presented in Tables 13 and 14.

Table 13. Simulation Results for Models 7 & 8 based on 100 realizations (Usual)

Method	ML2	MC+	MC0	%PS
$n = 72$ (Big and small effects)				
True	**0**	**7129**	**0**	**1**
Ridge	2.7303	10.05	0	0
Lasso	2.9437	7.38	0	0
Elastic Net	2.8957	8.56	0	0
Adaptive Lasso	3.284627	5.88	0	0
Adaptive Elastic Net	2.9253	8.53	0	0
SCAD	3.6833	9.91	0	0.01
$n = 72$ (Big, medium, and small effects)				
True	**0**	**7129**	**0**	**1**
Ridge	2.6420	10.93	0	0
Lasso	2.5961	8.33	0	0
Elastic Net	2.7088	9.31	0	0
Adaptive Lasso	3.3022	6.6	0	0
Adaptive Elastic Net	2.4906	9.32	0	0
SCAD	3.9524	10.98	0	0

Table 14. Simulation Results for Models 7 & 8 based on 100 realizations (**Relaxed**)

Method	ML2	MC+	MC0	%PS
$n = 72$ (Big and small effects)				
True	**0**	**5**	**7187**	**1**
Ridge	2.7238	4.99	0	0
Lasso	2.9403	4.85	7182.25	0
Elastic Net	2.8912	4.96	7079.98	0
Adaptive Lasso	3.2820	4.66	7184.82	0.07
Adaptive Elastic Net	2.9203	4.95	7079.98	0
SCAD	3.6812	4.89	0	0
$n = 72$ (Big, small, and medium effects)				
True	**0**	**7**	**7185**	**1**
Ridge	2.6443	6.51	0	0
Lasso	2.5976	6.04	7179.26	0
Elastic Net	2.7102	6.22	7177.81	0
Adaptive Lasso	3.3031	5.48	7181.18	0
Adaptive Elastic Net	2.4925	6.21	7177.81	0
SCAD	3.9485	6.46	0	0

5.3.4 Discussion

Several interesting observations can be made:

1. The performance of the estimators in the relaxed sense is much better than in the usual sense, except the performance of the SCAD in regard to %PS as before.

2. The performance of the estimators seem to be worse when there are medium effects along with big and small effects in the model.

3. The adaptive Lasso performs the best in the relaxed sense when there are only big and small effects in the model.

6.0 A BOOTSTRAP METHOD

6.1 SIGN CONSISTENCY

In the simulations in the last two chapters, the Lasso, Elastic Net, Adaptive Lasso, and Adaptive Elastic Net do well with respect to the measures ML2, MC+, and MC0 in the relaxed sense, but have a poor performance in regard to %PS (sign consistency). In other words, they have parameter estimation consistency, and can identify the big parameters in the long run, but if we implement them for one time as in a real application, they often miss some variables with big parameters or include some variables with small parameters in the model. This can be illustrated by the following boxplots of the parameter estimates. The dotted lines indicate the value of the true parameters.

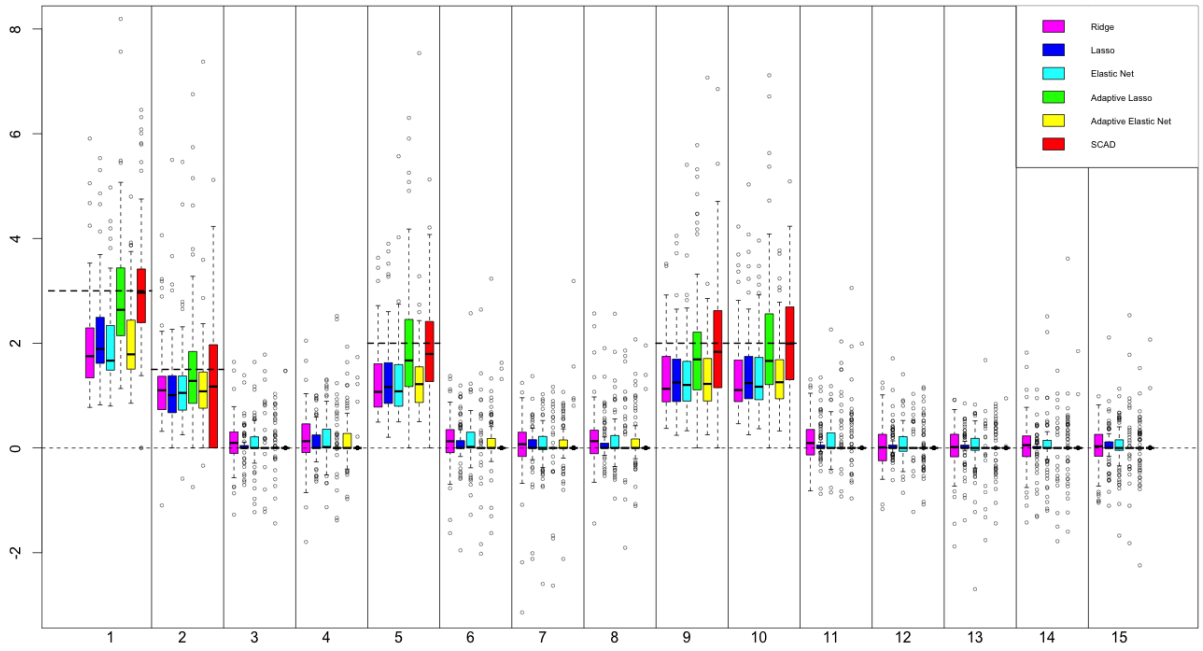

Figure 2: Boxplots of the Parameter Estimates of the First 15 Variables in Model 6 ($n = 80, p = 8,000 \rho = .5$)

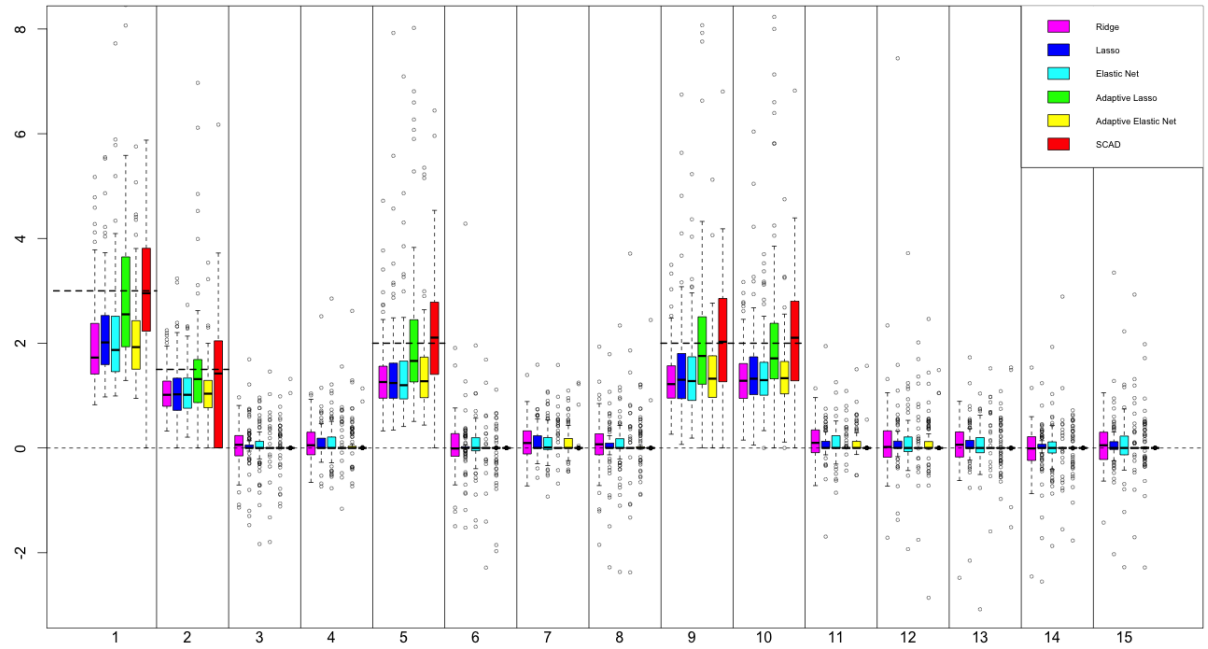

Figure 3: Boxplots of the Parameter Estimates of the First 15 Variables in Model 6 ($n = 80, p = 8,000, \rho = .3$)

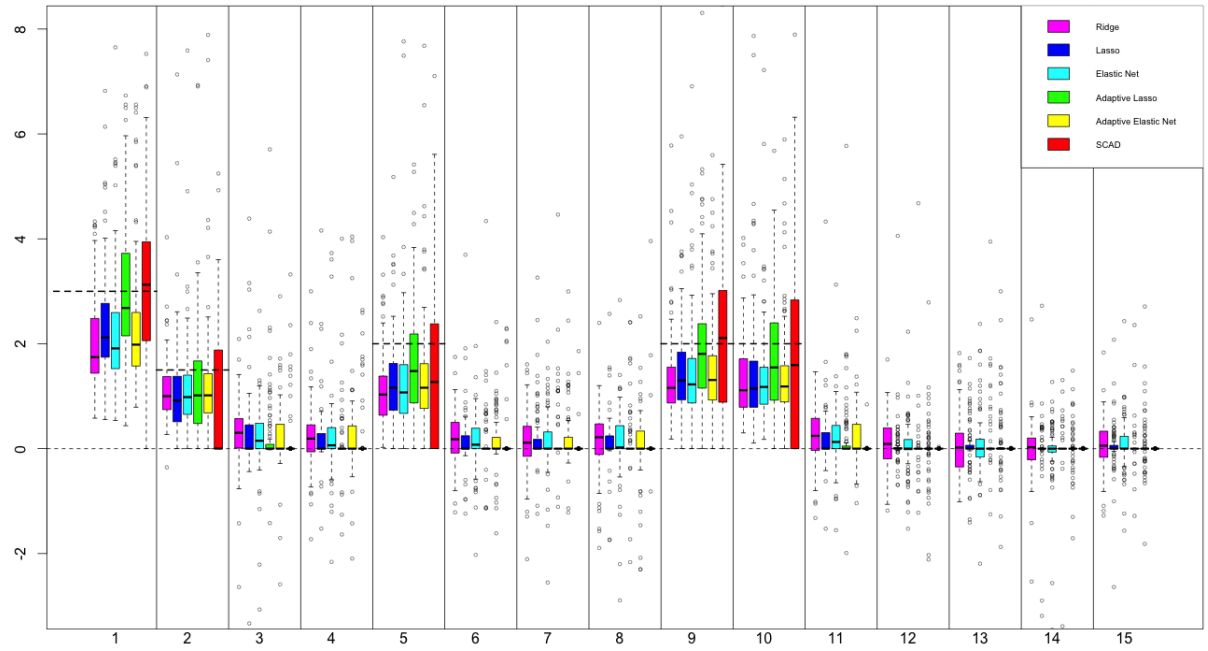

Figure 4: Boxplots of the Parameter Estimates of the First 15 Variables in Model 6 ($n = 80, p = 8,000, \rho = .7$)

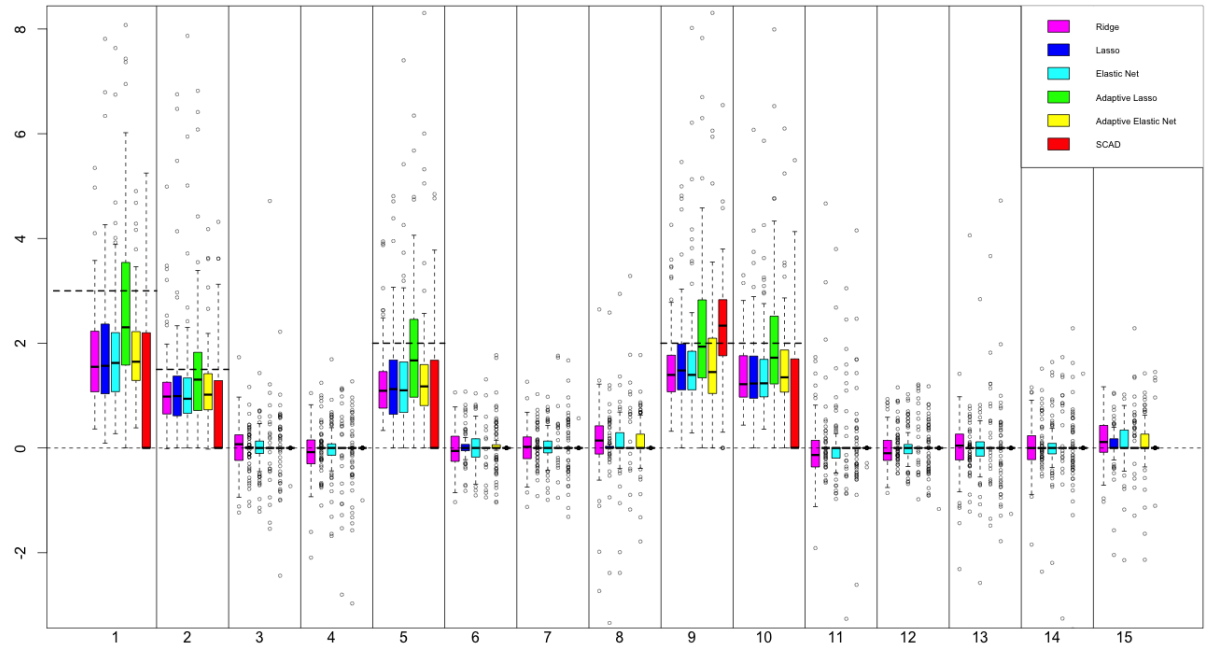

Figure 5: Boxplots of the Parameter Estimates of the First 15 Variables in Model 7 ($n = 72, p = 7,192$)

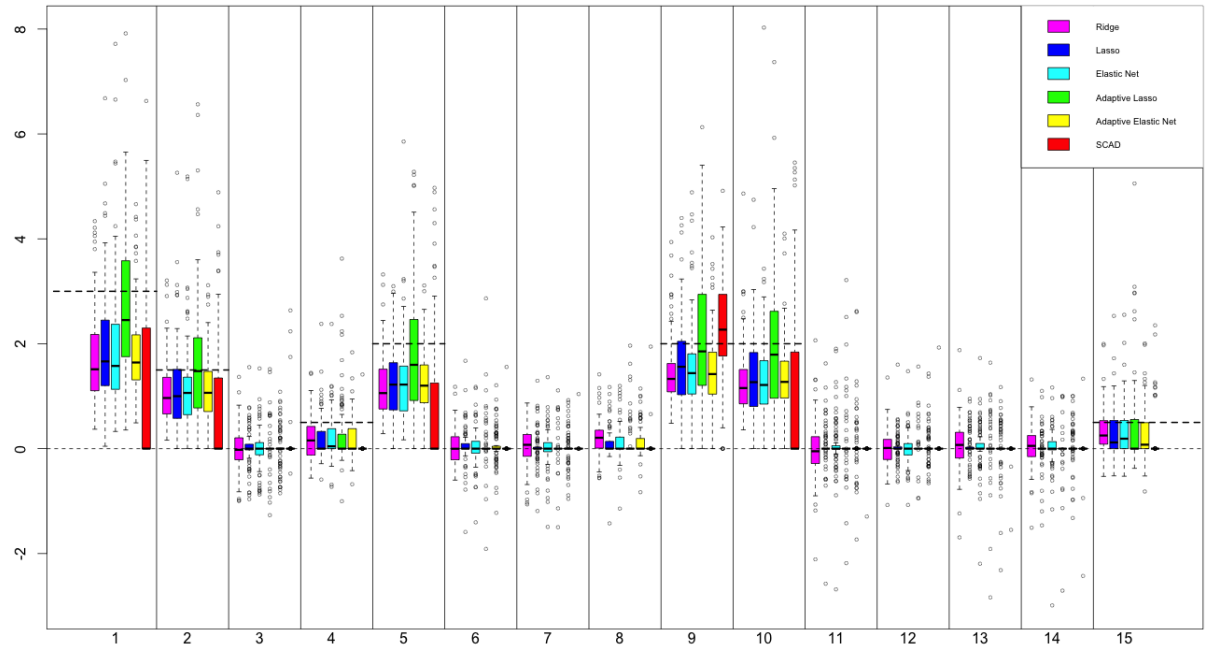

Figure 6: Boxplots of the Parameter Estimates of the First 15 Variables in Model 8 ($n = 72, p = 7,192$)

The patterns are obvious if we repeat the procedure for many times. The property of sign consistency can be greatly improved by repeating the procedure and taking the "average". This gives rise to a bootstrap method.

6.2 A BOOTSTRAP METHOD

The basic idea of the bootstrap method is to repeat the procedure on several bootstrap samples, say 100, and take advantage of the good "average" performance of the penalized estimators. We can look at the distribution of the votes of the penalized estimates based on different bootstrap samples, and only include those variables whose votes are greater than a threshold. We can set the threshold at where the biggest jump of the votes is. We can also use a majority vote idea and only include the variables who have more than 50% of the votes.

The bootstrap method can greatly improve the variable selection performance and reduce the false discovery rates. It can also give a sense of the variability in the parameter estimates. Yet, it is very computationally expensive, and it is hard to set the threshold and prove its asymptotic properties.

We simulate a single data set of size 100 for each of Models 6, 7, and 8. Then we generate 100 bootstrap samples for each of the data sets. The distribution of the votes of the estimates of the bootstrap samples are shown below. The dotted lines indicate the 50% thresholds. The SCAD tends to pick out a lot of variables with small parameters, thus, it is not good to use for the purpose of selecting variables. Hence, it is not included in the plots.

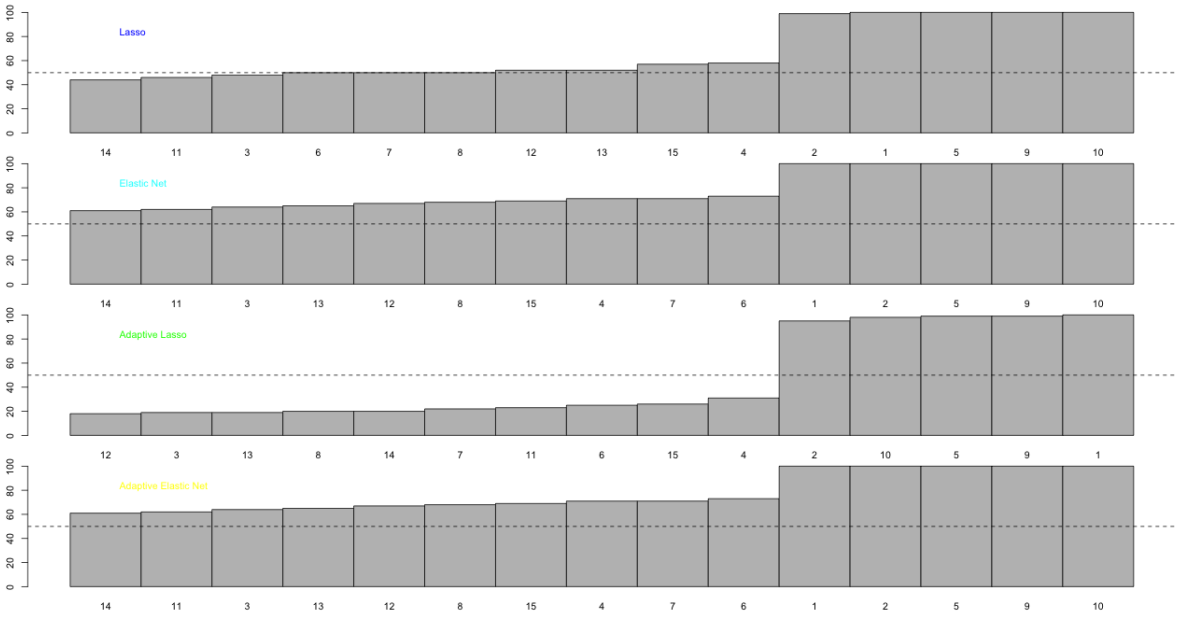

Figure 7: Distribution of the Votes of the First 15 Variables in Model 6 ($n = 80, p = 8,000, \rho = .5, \boldsymbol{\beta}^* = (\mathbf{3}, \mathbf{1.5}, \frac{5}{n}, \frac{5}{n}, \mathbf{2}, \frac{5}{n}, \frac{5}{n}, \frac{5}{n}, \mathbf{2}, \mathbf{2}, \frac{5}{n}, \cdots, \frac{5}{n})^T$)

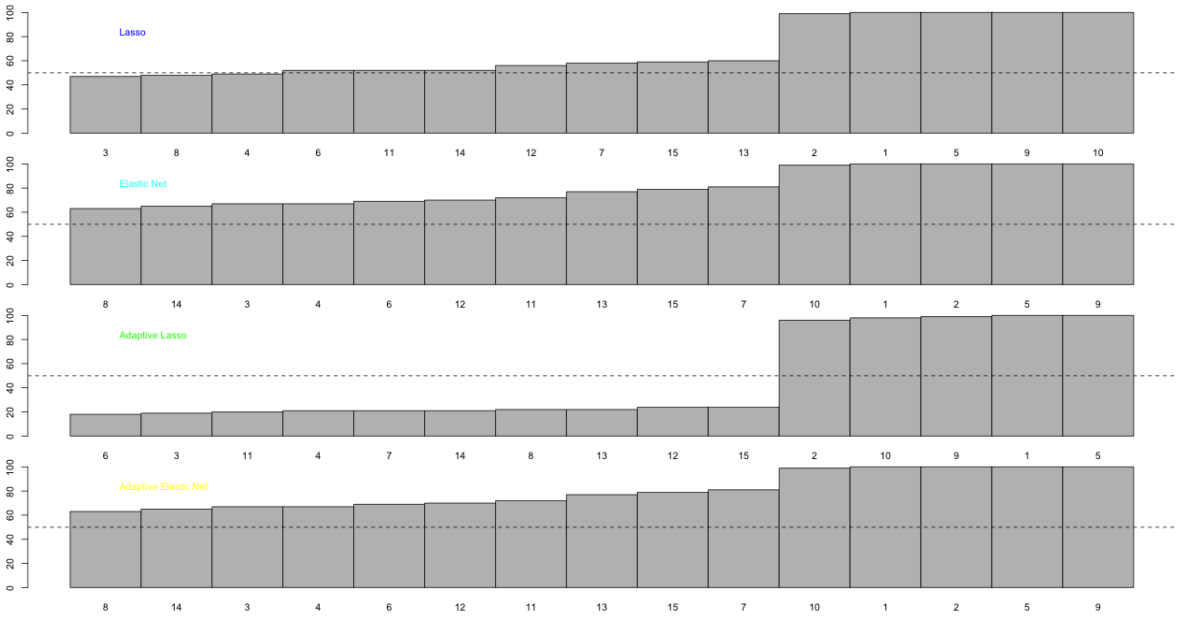

Figure 8: Distribution of the Votes of the First 15 Variables in Model 6 ($n = 80, p = 8,000, \rho = .3, \boldsymbol{\beta}^* = (\mathbf{3}, \mathbf{1.5}, \frac{5}{n}, \frac{5}{n}, \mathbf{2}, \frac{5}{n}, \frac{5}{n}, \frac{5}{n}, \mathbf{2}, \mathbf{2}, \frac{5}{n}, \cdots, \frac{5}{n})^T$)

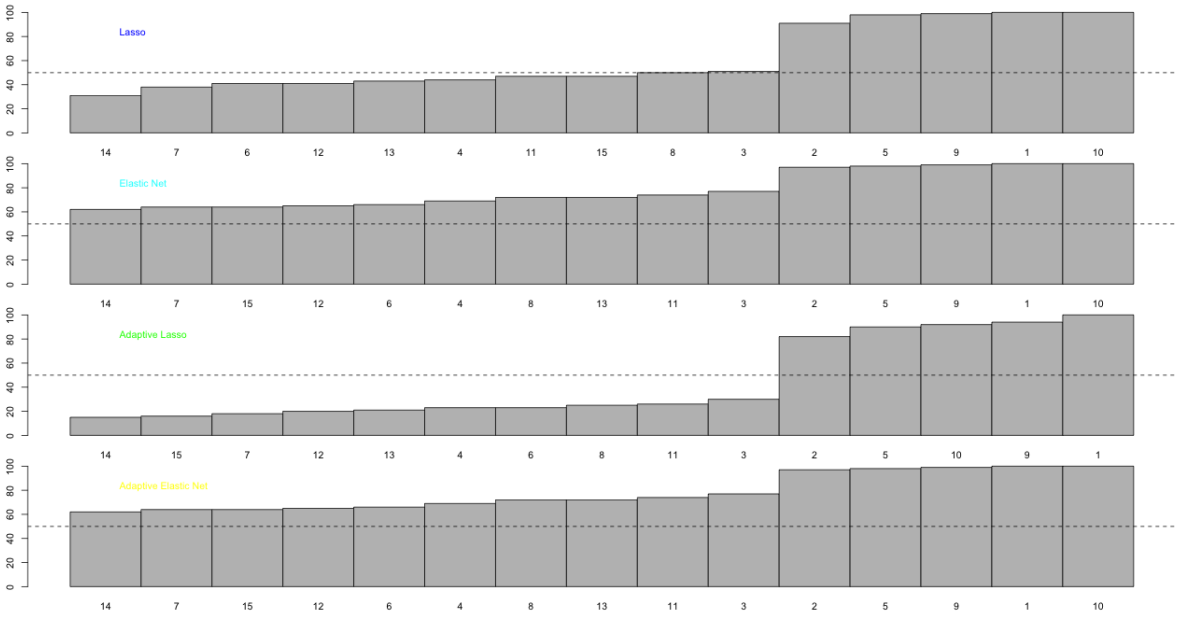

Figure 9: Distribution of the Votes of the First 15 Variables in Model 6 ($n = 80, p = 8,000, \rho = .7, \boldsymbol{\beta}^* = (\mathbf{3}, \mathbf{1.5}, \frac{5}{n}, \frac{5}{n}, \mathbf{2}, \frac{5}{n}, \frac{5}{n}, \frac{5}{n}, \mathbf{2}, \mathbf{2}, \frac{5}{n}, \cdots, \frac{5}{n})^T$)

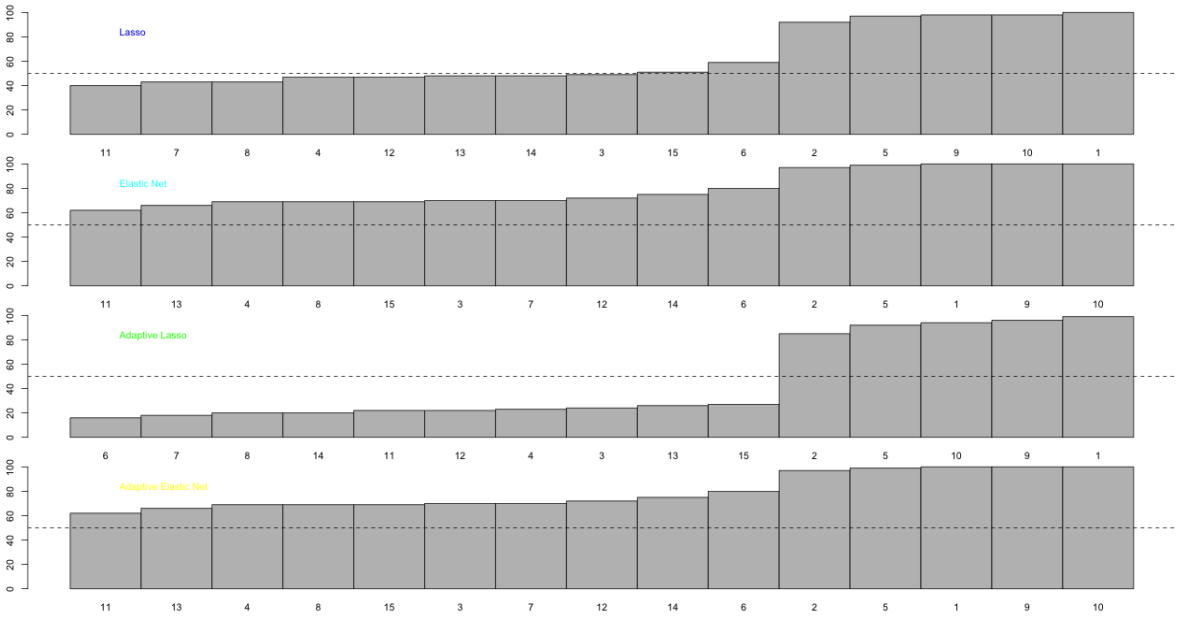

Figure 10: Distribution of the Votes of the First 15 Variables in Model 7 ($n = 72, p = 7,192, \boldsymbol{\beta}^* = (\mathbf{3}, \mathbf{1.5}, \frac{5}{n}, \frac{5}{n}, \mathbf{2}, \frac{5}{n}, \frac{5}{n}, \frac{5}{n}, \mathbf{2}, \mathbf{2}, \frac{5}{n}, \cdots, \frac{5}{n})^T$)

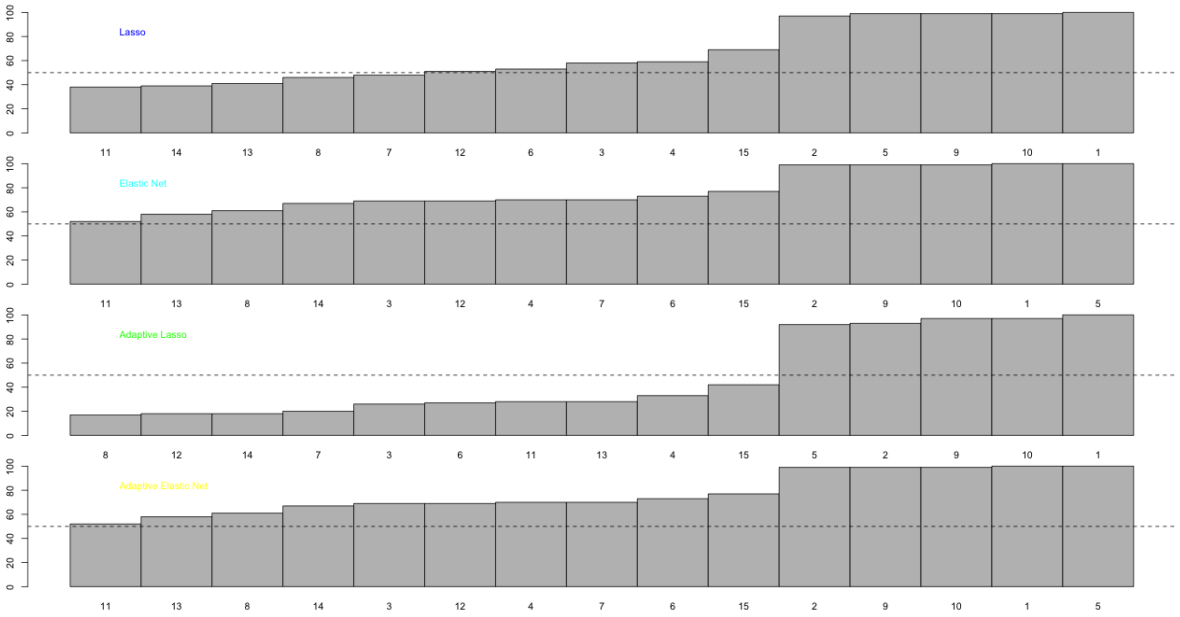

Figure 11: Distribution of the Votes of the First 15 Variables in Model 8 ($n = 72, p = 7,192, \boldsymbol{\beta}^* = (\mathbf{3}, \mathbf{1.5}, \frac{5}{n}, \mathbf{.5}, \mathbf{2}, \frac{5}{n}, \frac{5}{n}, \frac{5}{n}, \mathbf{2}, \mathbf{2}, \frac{5}{n}, \frac{5}{n}, \frac{5}{n}, \frac{5}{n}, \mathbf{.5}, \frac{5}{n}, \cdots, \frac{5}{n})^T$)

In the distributions of the votes of Models 6 and 7 where there are only big and small effects, there are clear jumps between the votes of the big 5 variables and the votes of the rest small variables. Yet, the threshold is at about 40% for the adaptive Lasso, about 60% for the Lasso, and about 80% for the elastic net and the adaptive elastic net. The adaptive Lasso has the biggest jump and lowest threshold. Variables $X1, X2, X5, X9$, and $X10$ which have big parameters are the only variables included in the final model if right thresholds are used.

In the distributions of the votes of Model 8 where there are medium effects along with big and small effects, the jumps between the votes of the big 5 variables and the votes of the remaining variables are clear but smaller than in models 6 and 7. There is no jump between the votes of the two variables who have medium effects and the votes of the rest variables who have small effects. It is much harder to distinguish the variables who have medium effects from the variables who have small effects than to distinguish the variables who have big effects from the variables who have medium or small effects. The threshold is at about 50% for the adaptive Lasso, about 70% for the Lasso, and about 80% for the elastic net and the adaptive elastic net. The adaptive Lasso has the biggest jump and lowest threshold. Variables $X1, X2, X5, X9$, and $X10$ which have big parameters are always included in the final model if right thresholds are used.

We may also build confidence intervals for the selected model. For example, we can use the variables that are selected by more than 50% of the penalized estimates of the bootstrap samples as the upper bound, and use the variables that are selected by all of the penalized estimates of the bootstrap samples as the lower bound. It is hard to determine the confidence level. More work needs to be done.

6.3 APPLICATION TO GENE EXPRESSION DATA

The data set is described in the last chapter. We apply the bootstrap method to the gene expression data with $n = 72, p = 7192$.

6.3.1 Results

Below are the boxplots of parameter estimates of 15 variables and the distribution of the votes of the same 15 variables.

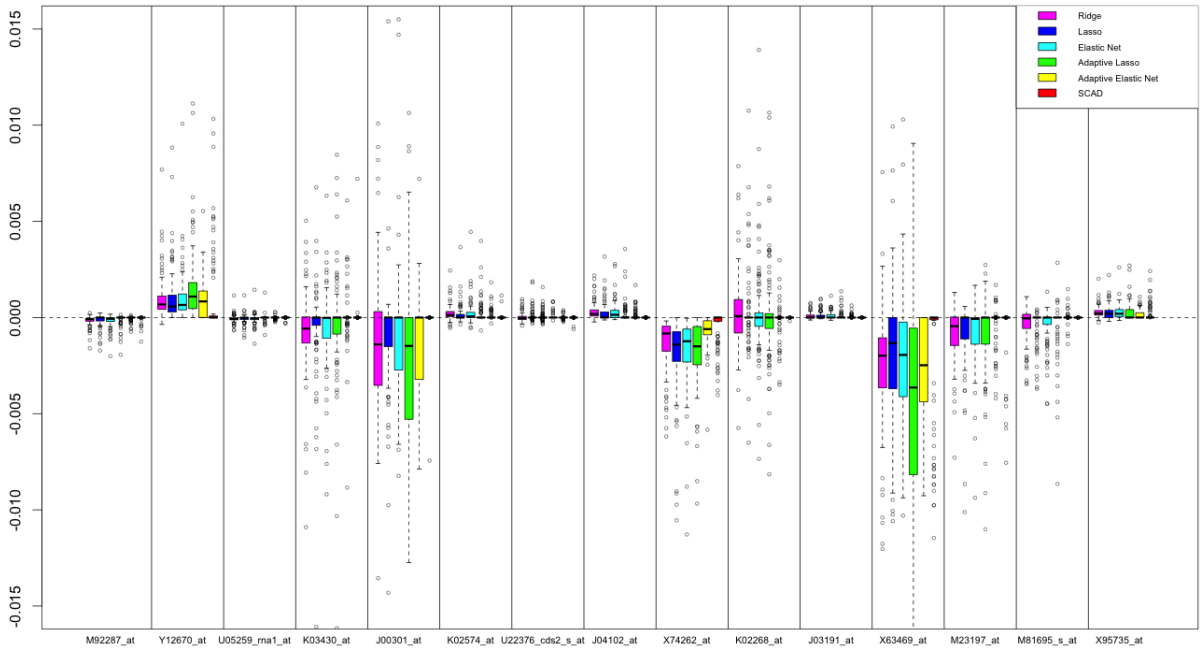

Figure 12: Boxplots of the Parameter Estimates of 15 Variables ($n = 72, p = 7,192$)

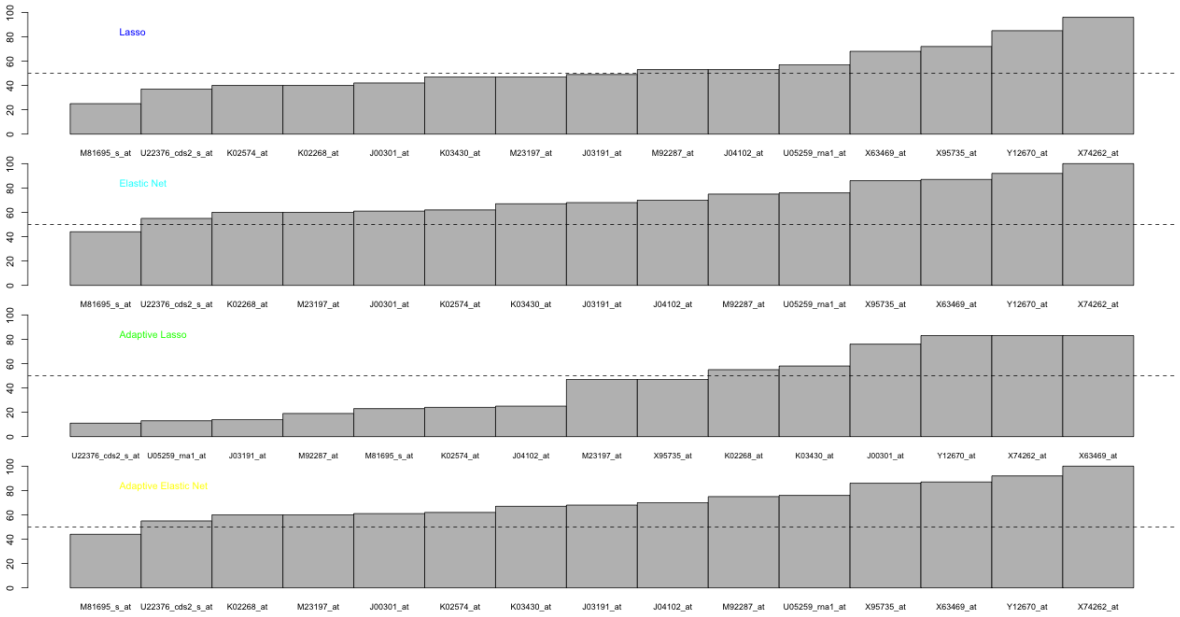

Figure 13: Distribution of the Votes of the Same 15 Variables ($n = 72, p = 7,192$)

6.3.2 Discussion and conclusion

Only the adaptive Lasso has a clear jump. And the adaptive Lasso almost always performs the best in our simulations. We then choose the adaptive Lasso as the estimator, and choose the threshold at 40% because that is where the biggest jump is. We end up with 43 variables.

Golub et al. (1999) selected 50 variables. There are 28 variables in common and many of them have been proven to be highly instructive in cancer classification. For example, CD33 encodes cell surface proteins for which monoclonal antibodies have been demonstrated to be useful in distinguishing lymphoid from myeloid lineage cells. The leptin receptor, originally identified through its role in weight regulation, showed high relative expression in AML. The leptin receptor was demonstrated to have anti-apoptotic function in hematopoietic cells. Similarly, the zyxin gene has been shown to encode a LIM domain protein important in cell adhesion in fibroblasts, but a role in hematopoiesis has not been reported. And some of the genes encode proteins critical for chromatin remodeling (RbAp48), and transcription (TFIIEβ).

The bootstrap method is very powerful in selecting the right set of variables. It can greatly improve variable selection performance and reduce false discovery rate. On the other hand, it is

very computationally expensive. More work is needed to select the right threshold and perhaps to give confidence intervals for the selected model.

7.0 FUTURE WORK

The main goal of this study is to explore the variable selection and parameter estimation properties of penalized likelihood estimators for GLMs in high-dimensional settings. We focused on the logistic model; we are interested in extending our work to the multi-logit, Poisson, and especially Cox's proportional hazards model.

We will further establish the Bootstrap method and study its asymptotic properties. We also will work on theoretical grounds to quantify the term non-influential and study how mid-size parameters may influence the performances of the penalized methods.

We will investigate how to determine the penalty parameter. Cross validation is computationally expensive, so we plan to study the BIC method. We will also study the variable selection and parameter estimation properties of the estimators whose penalty parameter is chosen by data-driven methods.

Further, we will explore the cases in which the sample size n and dimensionality p are both extremely large. Finite sample properties of the penalized likelihood estimators will also be examined.

We will also investigate whether adaptive weights can also be placed on the L_2 penalty, and will study its variable selection and parameter estimation properties. Other ways of incorporating information in the penalty will also be explored, along with their variable selection and parameter estimation properties. For example, penalties that make use of the grouping information will be explored; a penalty could be partly group Lasso penalty for dummy coding categorical variables.

Moreover, the Cox's proportional hazards model is of great interest and will be studied in the future because of genomic studies could well relate to survival.

BIBLIOGRAPHY

Breiman, L. (1995), "Better subset regression using the nonnegative garrote," *Technometrics*, 37, 373–384.

Candes, E. and Tao, T. (2007), "The Dantzig selector: statistical estimation when p is much larger than n," *Annals of Statistics*, 2313–2351.

Cawley, G. C. and Talbot, N. L. (2010), "On over-fitting in model selection and subsequent selection bias in performance evaluation," *Journal of Machine Learning Research*, 99, 2079–2107.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004), "Least angle regression," *Annals of Statistics*, 32, 407–499.

Fan, J., Fan, Y., and Lv, J. (2008), "High dimensional covariance matrix estimation using a factor model," *Journal of Econometrics*, 147, 186–197.

Fan, J. and Li, R. (2001), "Variable selection via nonconcave penalized likelihood and its oracle properties," *Journal of the American Statistical Association*, 96, 1348–1360.

Fan, J. and Li, R. (2006), "Statistical challenges with high dimensionality: Feature selection in knowledge discovery," *arXiv preprint math/0602133*.

Fan, J. and Lv, J. (2008), "Sure independence screening for ultrahigh dimensional feature space," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70, 849–911.

Fan, J. and Lv, J. (2011), "Nonconcave Penalized Likelihood With NP-Dimensionality," *Information Theory, IEEE Transactions on*, 57, 5467–5484.

Fan, J. and Peng, H. (2004), "Nonconcave penalized likelihood with a diverging number of parameters," *Annals of Statistics*, 32, 928–961.

Frank, L. E. and Friedman, J. H. (1993), "A statistical view of some chemometrics regression tools," *Technometrics*, 35, 109–135.

Friedman, J., Hastie, T., and Tibshirani, R. (2010), "Regularization paths for generalized linear models via coordinate descent," *Journal of Statistical Software*, 33, 1.

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., et al. (1999), "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring," *Science*, 286, 531–537.

Hastie, T., Tibshirani, R., and Friedman, J. J. H. (2001), *The elements of statistical learning*, vol. 1, Springer New York.

Hoerl, A. E. and Kennard, R. W. (1970), "Ridge regression: Biased estimation for nonorthogonal problems," *Technometrics*, 12, 55–67.

Huang, J., Horowitz, J. L., and Ma, S. (2008a), "Asymptotic properties of bridge estimators in sparse high-dimensional regression models," *Annals of Statistics*, 36, 587–613.

Huang, J., Ma, S., and Zhang, C.-H. (2008b), "Adaptive Lasso for sparse high-dimensional regression models," *Statistica Sinica*, 18, 1603.

Knight, K. and Fu, W. (2000), "Asymptotics for lasso-type estimators," *Annals of Statistics*, 1356–1378.

Leeb, H. and Pötscher, B. M. (2005), "Model selection and inference: Facts and fiction," *Econometric Theory*, 21, 21–59.

Leeb, H. and Pötscher, B. M. (2008), "Sparse estimators and the oracle property, or the return of Hodges' estimator," *Journal of Econometrics*, 142, 201–211.

Leng, C., Lin, Y., and Wahba, G. (2006), "A note on the lasso and related procedures in model selection," *Statistica Sinica*, 16, 1273.

MacCullagh, P. and Nelder, J. A. (1989), *Generalized linear models*, vol. 37, CRC press.

Meier, L., Van De Geer, S., and Bühlmann, P. (2008), "The group lasso for logistic regression," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70, 53–71.

Meinshausen, N. and Bühlmann, P. (2006), "High-dimensional graphs and variable selection with the lasso," *Annals of Statistics*, 34, 1436–1462.

Qian, G. and Wu, Y. (2006), "Strong limit theorems on model selection in generalized linear regression with binomial responses," *Statistica Sinica*, 16, 1335.

Rosset, S. and Zhu, J. (2007), "Piecewise linear regularized solution paths," *Annals of Statistics*, 1012–1030.

Tibshirani, R. (1996), "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, 267–288.

Tibshirani, R., Bien, J., Friedman, J., Hastie, T., Simon, N., Taylor, J., and Tibshirani, R. J. (2012), "Strong rules for discarding predictors in lasso-type problems," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74, 245–266.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., and Knight, K. (2005), "Sparsity and smoothness via the fused lasso," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67, 91–108.

Tibshirani, R. J. (2011), *The solution path of the generalized lasso*, Stanford University.

Tibshirani, R. J. and Taylor, J. (2012), "Degrees of freedom in lasso problems," *Annals of Statistics*, 40, 1198–1232.

Van de Geer, S. A. (2008), "High-dimensional generalized linear models and the Lasso," *Annals of Statistics*, 36, 614–645.

Yuan, M. and Lin, Y. (2006), "Model selection and estimation in regression with grouped variables," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68, 49–67.

Yuan, M. and Lin, Y. (2007), "On the non-negative garrotte estimator," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69, 143–161.

Zhao, M., Batista, A., Cunningham, J. P., Chestek, C., Rivera-Alvidrez, Z., Kalmar, R., Ryu, S., Shenoy, K., and Iyengar, S. (2012), "An L 1-regularized logistic model for detecting short-term neuronal interactions," *Journal of Computational Neuroscience*, 32, 479–497.

Zhao, P. and Yu, B. (2006), "On model selection consistency of Lasso," *Journal of Machine Learning Research*, 7, 2541–2563.

Zhou, S., van de Geer, S., and Bühlmann, P. (2009), "Adaptive Lasso for high dimensional regression and Gaussian graphical modeling," *arXiv preprint arXiv:0903.2515*.

Zou, H. (2006), "The adaptive lasso and its oracle properties," *Journal of the American Statistical Association*, 101, 1418–1429.

Zou, H. and Hastie, T. (2005), "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67, 301–320.

Zou, H. and Zhang, H. H. (2009), "On the adaptive elastic-net with a diverging number of parameters," *Annals of Statistics*, 37, 1733.